Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning

Francis Bach
INRIA - Sierra Project-Team
Ecole Normale Supérieure
Paris, France
francis.bach@inria.fr

Eric Moulines
LTCI
Telecom ParisTech
Paris, France
eric.moulines@enst.fr

July 12, 2011

Abstract

In this paper, we consider the minimization of a convex objective function defined on a Hilbert space, which is only available through unbiased estimates of its gradients. This problem includes standard machine learning algorithms such as kernel logistic regression and least-squares regression, and is commonly referred to as a stochastic approximation problem in the operations research community. We provide a non-asymptotic analysis of the convergence of two well-known algorithms, stochastic gradient descent (a.k.a. Robbins-Monro algorithm) as well as a simple modification where iterates are averaged (a.k.a. Polyak-Ruppert averaging). Our analysis suggests that a learning rate proportional to the inverse of the number of iterations, while leading to the optimal convergence rate in the strongly convex case, is not robust to the lack of strong convexity or the setting of the proportionality constant. This situation is remedied when using slower decays together with averaging, robustly leading to the optimal rate of convergence. We illustrate our theoretical results with simulations on synthetic and standard datasets.

1 Introduction

The minimization of an objective function which is only available through unbiased estimates of the function values or its gradients is a key methodological problem in many disciplines. Its analysis has been attacked mainly in three communities: stochastic approximation [1, 2, 3, 4, 5, 6], optimization [7, 8], and machine learning [9, 10, 11, 12, 13, 14, 15]. The main algorithms which have emerged are stochastic gradient descent (a.k.a. Robbins-Monro algorithm), as well as a simple modification where iterates are averaged (a.k.a. Polyak-Ruppert averaging).

Traditional results from stochastic approximation rely on strong convexity and asymptotic analysis, but have made clear that a learning rate proportional to the inverse of the number of iterations, while leading to the optimal convergence rate in the strongly convex case, is not robust to the wrong setting of the proportionality constant. On the other hand, using slower decays together with averaging robustly leads to optimal convergence behavior (both in terms of rates and constants) [4, 5].

The analysis from the convex optimization and machine learning literatures however has focused on differences between strongly convex and non-strongly convex objectives, with learning rates and roles of averaging being different in these two cases [11, 12, 13, 14, 15].

A key desirable behavior of an optimization method is to be adaptive to the hardness of the problem, and thus one would like a single algorithm to work in all situations, favorable ones such as strongly convex functions and unfavorable ones such as non-strongly convex functions. In this paper, we unify the two types of analysis and show that (1) a learning rate proportional to the inverse of the number of iterations is not suitable because it is not robust to the setting of the proportionality constant and the lack of strong convexity, (2) the use of averaging with slower decays allows (close to) optimal rates in *all* situations.

More precisely, we make the following contributions:

- We provide a direct non-asymptotic analysis of stochastic gradient descent in a machine learning context (observations of real random functions defined on a Hilbert space) that includes kernel least-squares regression and logistic regression (see Section 2), with strong convexity assumptions (Section 3) and without (Section 4).
- We provide a non-asymptotic analysis of Polyak-Ruppert averaging [4, 5], with and without strong convexity (Sections 3.3 and 4.2). In particular, we show that slower decays of the learning rate, *together with averaging*, are crucial to *robustly* obtain fast convergence rates.
- We illustrate our theoretical results through experiments on synthetic and non-synthetic examples in Section 5.

Notation. We consider a Hilbert space $\mathcal{H}$ with a scalar product $\langle \cdot, \cdot \rangle$. We denote by $\| \cdot \|$ the associated norm and use the same notation for the operator norm on bounded linear operators from $\mathcal{H}$ to $\mathcal{H}$, defined as $\|A\| = \sup_{\|x\| \leqslant 1} \|Ax\|$ (if $\mathcal{H}$ is a Euclidean space, then $\|A\|$ is the largest singular value of A). We also use the notation "w.p.1" to mean "with probability one". We denote by $\mathbb{E}$ the expectation or conditional expectation with respect to the underlying probability space.

2 Problem set-up

We consider a sequence of *convex differentiable random* functions $(f_n)_{n \geqslant 1}$ from $\mathcal{H}$ to $\mathbb{R}$. We consider the following recursion, starting from $\theta_0 \in \mathcal{H}$:

$$\forall n \geqslant 1, \quad \theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1}), \tag{1}$$

where $(\gamma_n)_{n \geqslant 1}$ is a deterministic sequence of positive scalars, which we refer to as the *learning rate sequence*. The function f_n is assumed to be differentiable (see, e.g., [16] for definitions and properties of differentiability for functions defined on Hilbert spaces), and its gradient is an unbiased estimate of the gradient of a certain function f we wish to minimize:

(**H1**) Let $(\mathcal{F}_n)_{n \geqslant 0}$ be an increasing family of σ-fields. θ_0 is $\mathcal{F}_0$-measurable, and for each $\theta \in \mathcal{H}$, the random variable $f_n'(\theta)$ is square-integrable, $\mathcal{F}_n$-measurable and

$$\forall \theta \in \mathcal{H}, \quad \forall n \geqslant 1, \quad \mathbb{E}(f_n'(\theta) | \mathcal{F}_{n-1}) = f'(\theta), \text{ w.p.1.} \tag{2}$$

For an introduction to martingales, σ-fields, and conditional expectations, see, e.g., [17]. Note that depending whether $\mathcal{F}_0$ is a trivial σ-field or not, θ_0 may be random or not. Moreover, we could restrict Eq. (2) to be satisfied only for θ_{n-1} and θ^* (which is a global minimizer of f).

Given only the noisy gradients $f_n'(\theta_{n-1})$, the goal of stochastic approximation is to minimize the function f with respect to θ. Our assumptions include two usual situations, but also include many others (e.g., active learning):

– **Stochastic approximation**: in the so-called Robbins-Monro setting, for all $\theta \in \mathcal{H}$ and $n \geqslant 1$, $f_n(\theta)$ may be expressed as $f_n(\theta) = f(\theta) + \langle \varepsilon_n, \theta \rangle$, where $(\varepsilon_n)_{n \geqslant 1}$ is a square-integrable martingale difference (i.e., such that $\mathbb{E}(\varepsilon_n | \mathcal{F}_{n-1}) = 0$), which corresponds to a noisy observation $f'(\theta_{n-1}) + \varepsilon_n$ of the gradient $f'(\theta_{n-1})$.

– **Learning from i.i.d. observations**: for all $\theta \in \mathcal{H}$ and $n \geqslant 1$, $f_n(\theta) = \ell(\theta, z_n)$ where z_n is an i.i.d. sequence of observations in a measurable space $\mathcal{Z}$ and $\ell : \mathcal{H} \times \mathcal{Z}$ is a loss function. Then $f(\theta)$ is the generalization error of the predictor defined by θ. Classical examples are least-squares or logistic regression (linear or non-linear through kernel methods [18, 19]), where $f_n(\theta) = \frac{1}{2}(\langle x_n, \theta \rangle - y_n)^2$, or $f_n(\theta) = \log[1 + \exp(-y_n \langle x_n, \theta \rangle)]$, for $x_n \in \mathcal{H}$, and $y_n \in \mathbb{R}$ (or $\{-1, 1\}$ for logistic regression).

Throughout this paper, we assume that each function f_n is convex and *smooth*, following the traditional definition of smoothness from the optimization literature, i.e., Lipschitz-continuity of the gradients (see, e.g., [20]). However, we make two slightly different assumptions: **(H2)** where the function $\theta \mapsto \mathbb{E}(f_n'(\theta) | \mathcal{F}_{n-1})$ is Lipschitz-continuous in quadratic mean and a strengthening of this assumption, **(H2')** in which $\theta \mapsto f_n'(\theta)$ is almost surely Lipschitz-continuous.

(**H2**) For each $n \geqslant 1$, the function f_n is almost surely convex, differentiable, and:

$$\forall n \geqslant 1, \ \forall \theta_1, \theta_2 \in \mathcal{H}, \quad \mathbb{E}(\|f_n'(\theta_1) - f_n'(\theta_2)\|^2 | \mathcal{F}_{n-1}) \leqslant L^2 \|\theta_1 - \theta_2\|^2 , \quad \text{w.p.1.} \tag{3}$$

(**H2'**) For each $n \geqslant 1$, the function f_n is almost surely convex, differentiable with Lipschitz-continuous gradient f_n', with constant L, that is:

$$\forall n \geqslant 1, \ \forall \theta_1, \theta_2 \in \mathcal{H}, \quad \|f_n'(\theta_1) - f_n'(\theta_2)\| \leqslant L \|\theta_1 - \theta_2\| , \quad \text{w.p.1.} \tag{4}$$

If f_n is twice differentiable, this corresponds to having the operator norm of the Hessian operator of f_n bounded by L. For least-squares or logistic regression, if we assume that $(\mathbb{E}\|x_n\|^4)^{1/4} \leqslant R$ for all $n \in \mathbb{N}$, then we may take $L = R^2$ (or even $L = R^2/4$ for logistic regression) for assumption (**H2**), while for assumption **(H2')**, we need to have an almost sure bound $\|x_n\| \leqslant R$.

3 Strongly convex objectives

In this section, following [21], we make the additional assumption of strong convexity of f, but not of all functions f_n (see [20] for definitions and properties of such functions):

(**H3**) The function f is strongly convex with respect to the norm $\|\cdot\|$, with convexity constant $\mu > 0$. That is, for all $\theta_1, \theta_2 \in \mathcal{H}$, $f(\theta_1) \geqslant f(\theta_2) + \langle f'(\theta_2), \theta_1 - \theta_2 \rangle + \frac{\mu}{2} \|\theta_1 - \theta_2\|^2$.

Note that (**H3**) simply needs to be satisfied for $\theta_2 = \theta^*$ being the unique global minimizer of f (such that $f'(\theta^*) = 0$). In the context of machine learning (least-squares or logistic regression), assumption **(H3)** is satisfied as soon as $\frac{\mu}{2}\|\theta\|^2$ is used as an additional regularizer. For all strongly convex losses (e.g., least-squares), it is also satisfied as soon as the expectation $\mathbb{E}(x_n \otimes x_n)$ is invertible. Note that this implies that the problem is finite-dimensional, otherwise, the expectation is a compact covariance operator, and hence non-invertible (see, e.g., [22] for an introduction to covariance operators). For non-strongly convex losses such as the logistic loss, f can never be strongly convex unless we restrict the domain of θ (which we do in Section 3.2). Alternatively to restricting the domain, replacing the logistic loss $u \mapsto \log(1 + e^{-u})$ by $u \mapsto \log(1 + e^{-u}) + \varepsilon u^2/2$, for some small $\varepsilon > 0$, makes it strongly convex in low-dimensional settings.

By strong convexity of f, if we assume (**H3**), then f attains its global minimum at a unique vector $\theta^* \in \mathcal{H}$ such that $f'(\theta^*) = 0$. Moreover, we make the following assumption (in the context of stochastic approximation, it corresponds to $\mathbb{E}(\|\varepsilon_n\|^2|\mathcal{F}_{n-1}) \leqslant \sigma^2$):

(**H4**) There exists $\sigma^2 \in \mathbb{R}_+$ such that for all $n \geqslant 1$, $\mathbb{E}(\|f_n'(\theta^*)\|^2|\mathcal{F}_{n-1}) \leqslant \sigma^2$, w.p.1.

3.1 Stochastic gradient descent

Before stating our first theorem (see proof in the appendix), we introduce the following family of functions: $\varphi_\beta : \mathbb{R}_+ \setminus \{0\} \to \mathbb{R}$ given by:

$$\varphi_\beta(t) = \begin{cases} \frac{t^\beta - 1}{\beta} & \text{if } \beta \neq 0, \\ \log t & \text{if } \beta = 0. \end{cases}$$

The function $\beta \mapsto \varphi_\beta(t)$ is continuous for all $t > 0$. Moreover, for $\beta > 0$, $\varphi_\beta(t) < \frac{t^\beta}{\beta}$, while for $\beta < 0$, we have $\varphi_\beta(t) < \frac{1}{-\beta}$ (both with asymptotic equality when t is large).

Theorem 1 (Stochastic gradient descent, strong convexity) *Assume* (**H1**,**H2**,**H3**,**H4**). *Denote by* $\delta_n = \mathbb{E}\|\theta_n - \theta^*\|^2$, *where* $\theta_n \in \mathcal{H}$ *is the* n*-th iterate of the recursion in Eq. (1), with* $\gamma_n = Cn^{-\alpha}$. *We have, for* $\alpha \in [0, 1]$*:*

$$\delta_n \leqslant \begin{cases} 2\exp\left(4L^2C^2\varphi_{1-2\alpha}(n)\right)\exp\left(-\frac{\mu C}{4}n^{1-\alpha}\right)\left(\delta_0 + \frac{\sigma^2}{L^2}\right) + \frac{4C\sigma^2}{\mu n^\alpha}, & \text{if } \ 0 \leqslant \alpha < 1, \\ \frac{\exp(2L^2C^2)}{n^{\mu C}}\left(\delta_0 + \frac{\sigma^2}{L^2}\right) + 2\sigma^2C^2\frac{\varphi_{\mu C/2-1}(n)}{n^{\mu C/2}}, & \text{if } \ \alpha = 1. \end{cases} \tag{5}$$

Sketch of proof. Under our assumptions, it can be shown that (δ_n) satisfies the following recursion:

$$\delta_n \leqslant (1 - 2\mu\gamma_n + 2L^2\gamma_n^2)\delta_{n-1} + 2\sigma^2\gamma_n^2. \tag{6}$$

Note that it also appears in [3, Eq. (2)] under different assumptions. Using this recursion, we then derive bounds using classical techniques from stochastic approximation [2], but in a non-asymptotic way, by deriving explicit upper-bounds.

Related work. To the best of our knowledge, this non-asymptotic bound, which depends explicitly upon the parameters of the problem, is novel (see [1, Theorem 1, Electronic companion paper] for a simpler bound with no such explicit dependence). It shows in particular that there is convergence in quadratic mean for any $\alpha \in (0, 1]$. Previous results from the stochastic approximation literature have focused mainly on almost sure convergence of the sequence of iterates. Almost-sure convergence requires that $\alpha > 1/2$, with counter-examples for $\alpha < 1/2$ (see, e.g., [2] and references therein).

Bound on function values. The bounds above imply a corresponding a bound on the functions values. Indeed, under assumption **(H2)**, it may be shown that $\mathbb{E}[f(\theta_n) - f(\theta^*)] \leqslant \frac{L}{2}\delta_n$ (see proof in the appendix).

Tightness for quadratic functions. Since the deterministic recursion in Eq. (6) is an equality for quadratic functions f_n, the result in Eq. (5) is optimal (up to constants). Moreover, our results are consistent with the asymptotic results from [6].

Forgetting initial conditions. Bounds depend on the initial condition $\delta_0 = \mathbb{E}\left[\|\theta_0 - \theta^*\|^2\right]$ and the variance σ^2 of the noise term. The initial condition is forgotten sub-exponentially fast for $\alpha \in (0, 1)$, but not for $\alpha = 1$. For $\alpha < 1$, the asymptotic term in the bound is $\frac{4C\sigma^2}{\mu n^\alpha}$.

Behavior for $\alpha = 1$. For $\alpha = 1$, we have $\frac{\varphi_{\mu C/2-1}(n)}{n^{\mu C/2}} \leqslant \frac{1}{\mu C/2-1}\frac{1}{n}$ if $C\mu > 2$, $\frac{\varphi_{\mu C/2-1}(n)}{n^{\mu C/2}} = \frac{\log n}{n}$ if $C\mu = 2$ and $\frac{\varphi_{\mu C/2-1}(n)}{n^{\mu C/2}} \leqslant \frac{1}{1-\mu C/2}\frac{1}{n^{\mu C/2}}$ if $C\mu > 2$. Therefore, for $\alpha = 1$, the choice of C is critical,

as already noticed by [8]: too small C leads to convergence at arbitrarily small rate of the form $n^{-\mu C/2}$, while too large C leads to explosion due to the initial condition. This behavior is confirmed in simulations in Section 5.

Setting C too large. There is a potentially catastrophic term when C is chosen too large, i.e., $\exp\big(4L^2C^2\varphi_{1-2\alpha}(n)\big)$, which leads to an increasing bound when n is small. Note that for $\alpha < 1$, this catastrophic term is in front of a sub-exponentially decaying factor, so its effect is mitigated once the term in $n^{1-\alpha}$ takes over $\varphi_{1-2\alpha}(n)$, and the transient term stops increasing. Moreover, the asymptotic term is not involved in it (which is also observed in simulations in Section 5).

Minimax rate. Note finally, that the asymptotic convergence rate in $O(n^{-1})$ matches optimal asymptotic minimax rate for stochastic approximation [23].

3.2 Bounded gradients

In some cases such as logistic regression, we also have a uniform upper-bound on the gradients, i.e., we assume (note that in Theorem 2, this assumption replaces both **(H2)** *and* **(H4)**).

(**H5**) For each $n \geqslant 1$, almost surely, the function f_n if convex, differentiable and has gradients uniformly bounded by B on the ball of center 0 and radius D, i.e., for all $\theta \in \mathcal{H}$ and all $n > 0$, $\|\theta\| \leqslant D \Rightarrow \|f_n'(\theta)\| \leqslant B$.

Note that no function may be strongly convex and Lipschitz-continuous (i.e., with uniformly bounded gradients) over the entire Hilbert space $\mathcal{H}$. Indeed, if the function is μ-strongly convex and $\|\theta^*\| \leqslant D/2$, then (**H5**) implies that $B \geqslant \mu D/2$ (straightforward consequence of [20, Eq. (2.1.22)]). Moreover, if (**H2'**) is satisfied, then we may take $D = \|\theta^*\|$ and $B = LD$. The next theorem shows that with a slight modification of the recursion in Eq. (1), we get simpler bounds than the ones obtained in Theorem 1, obtaining a result which already appeared in a simplified form [8] (see proof in the appendix):

Theorem 2 (Stochastic gradient descent, strong convexity, bounded gradients) *Assume* (**H1**,**H3**,**H5**). *Denote* $\delta_n = \mathbb{E}\left[\|\theta_n - \theta^*\|^2\right]$, *where* $\theta_n \in \mathcal{H}$ *is the* n*-th iterate of the following recursion:*

$$\forall n \geqslant 1, \quad \theta_n = \Pi_D[\theta_{n-1} - \gamma_n f_n'(\theta_{n-1})], \tag{7}$$

where Π_D *is the orthogonal projection operator on the ball* $\{\theta : \|\theta\| \leqslant D\}$. *Assume* $\|\theta^*\| \leqslant D$. *If* $\gamma_n = Cn^{-\alpha}$, *we have, for* $\alpha \in [0,1]$*:*

$$\delta_n \leqslant \begin{cases} \big(\delta_0 + B^2C^2\varphi_{1-2\alpha}(n)\big)\exp\left(-\frac{\mu C}{2}n^{1-\alpha}\right) + \frac{2B^2C^2}{\mu n^\alpha}, & \text{if } \alpha \in [0,1)\,; \\ \delta_0 n^{-\mu C} + 2B^2C^2 n^{-\mu C}\varphi_{\mu C-1}(n), & \text{if } \alpha = 1\,. \end{cases} \tag{8}$$

The proof follows the same lines than for Theorem 1, but with the deterministic recursion $\delta_n \leqslant (1 - 2\mu\gamma_n)\delta_{n-1} + B^2\gamma_n^2$. Note that we obtain the same asymptotic terms than for Theorem 1 (but B replaces σ). Moreover, the bound is simpler (no explosive multiplicative factors), but it requires to know D in advance, while Theorem 1 does not.

3.3 Polyak-Ruppert averaging

We now consider $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k$ and, following [4, 5], we make extra assumptions regarding the smoothness of each f_n and the fourth-order moment of the driving noise:

(**H7**) For each $n \geqslant 1$, the function f_n is almost surely twice differentiable with Lipschitz-continuous Hessian operator f_n'', with Lipschitz constant M. That is, for all $\theta_1, \theta_2 \in \mathcal{H}$ and for all $n \geqslant 1$, $\|f_n''(\theta_1) - f_n''(\theta_2)\| \leqslant M\|\theta_1 - \theta_2\|$, where $\|\cdot\|$ is the operator norm.

Note that (**H7**) needs only to be satisfied for $\theta_2 = \theta^*$. For least-square regression, we have $M = 0$, while for logistic regression, we have $M = R^3/4$. Note that it may be possible to use recent results in self-concordance analysis to link M and μ [24].

(**H8**) There exists $\tau \in \mathbb{R}_+$, such that for each $n \geqslant 1$, $\mathbb{E}(\|f_n'(\theta^*)\|^4 | \mathcal{F}_{n-1}) \leqslant \tau^4$ almost surely. Moreover, there exists a nonnegative self-adjoint operator Σ such that for all n, $\mathbb{E}(f_n'(\theta^*) \otimes f_n'(\theta^*) | \mathcal{F}_{n-1}) \preccurlyeq \Sigma$ almost-surely.

The operator Σ (which always exists as soon as τ is finite) is here to characterize precisely the variance term, which will be invariant of the learning rate sequence (γ_n), as we now show:

Theorem 3 (Averaging, strong convexity) *Assume* (**H1**, **H2'**, **H3**, **H4**, **H7**, **H8**). *Then, for* $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1} \theta_k$ *and* $\alpha \in (0,1)$, *we have:*

$$\begin{aligned}\left(\mathbb{E}\|\bar{\theta}_n - \theta^*\|^2\right)^{1/2} \leqslant & \frac{\left[\operatorname{tr} f''(\theta^*)^{-1}\Sigma f''(\theta^*)^{-1}\right]^{1/2}}{\sqrt{n}} + \frac{6\sigma}{\mu C^{1/2}}\frac{1}{n^{1-\alpha/2}} + \frac{MC\tau^2}{2\mu^{3/2}}(1+(\mu C)^{1/2})\frac{\varphi_{1-\alpha}(n)}{n} \\ & + \frac{4LC^{1/2}}{\mu}\frac{\varphi_{1-\alpha}(n)^{1/2}}{n} + \frac{8A}{n\mu^{1/2}}\Big(\frac{1}{C} + L\Big)\Big(\delta_0 + \frac{\sigma^2}{L^2}\Big)^{1/2} \\ & + \frac{5MC^{1/2}\tau}{2n\mu}A\exp\left(24L^4C^4\right)\Big(\delta_0 + \frac{\mu\mathbb{E}\left[\|\theta_0 - \theta^*\|^4\right]}{20C\tau^2} + 2\tau^2C^3\mu + 8\tau^2C^2\Big)^{1/2}, \end{aligned} \tag{9}$$

where A *is a constant that depends only on* μ, C, L *and* α.

Sketch of proof. Following [4], we start from Eq. (1), write it as $f_n'(\theta_{n-1}) = \frac{1}{\gamma_n}(\theta_{n-1} - \theta_n)$, and notice that (a) $f_n'(\theta_{n-1}) \approx f_n'(\theta^*) + f''(\theta^*)(\theta_{n-1} - \theta^*)$, (b) $f_n'(\theta^*)$ has zero mean and behaves like an i.i.d. sequence, and (c) $\frac{1}{n}\sum_{k=1}^n \frac{1}{\gamma_k}(\theta_{k-1} - \theta_k)$ turns out to be negligible owing to a summation by parts and to the bound obtained in Theorem 1. This implies that $\bar{\theta}_n - \theta^*$ behaves like $-\frac{1}{n}\sum_{k=1}^n f''(\theta^*)^{-1} f_k'(\theta^*)$. Note that we obtain a bound on the *root* mean square error.

Forgetting initial conditions. There is no sub-exponential forgetting of initial conditions, but rather a decay at rate $O(n^{-2})$ (last two lines in Eq. (9)). This is a known problem which may slow down the convergence, a common practice being to start averaging after a certain number of iterations [2]. Moreover, the constant A may be large when LC is large, thus the catastrophic terms are more problematic than for stochastic gradient descent, because they do not appear in front of sub-exponentially decaying terms (see appendix). This suggests to take CL small.

Asymptotically leading term. When $M > 0$ and $\alpha > 1/2$, the asymptotic term for δ_n is independent of (γ_n) and of order $O(n^{-1})$. Thus, averaging allows to get from the slow rate $O(n^{-\alpha})$ to the optimal rate $O(n^{-1})$. The next two leading terms (in the first line) have order $O(n^{\alpha-2})$ and $O(n^{-2\alpha})$, suggesting the setting $\alpha = 2/3$ to make them equal. When $M = 0$ (quadratic functions), the leading term has rate $O(n^{-1})$ for all $\alpha \in (0,1)$ (with then a contribution of the first term in the second line).

Case $\alpha = 1$. We get a simpler bound by directly averaging the bound in Theorem 1, which leads to an unchanged rate of n^{-1}, i.e., averaging is not key for $\alpha = 1$, and does not solve the robustness problem related to the choice of C or the lack of strong convexity.

Leading term independent of (γ_n). The term in $O(n^{-1})$ does not depend on γ_n. Moreover, as noticed in the stochastic approximation literature [4], in the context of learning from i.i.d. observations, this is exactly the Cramer-Rao bound (see, e.g., [25]), and thus the leading term is

asymptotically optimal. Note that no explicit Hessian inversion has been performed to achieve this bound.

Relationship with prior work on online learning. There is no clear way of adding a bounded gradient assumption in the general case $\alpha \in (0,1)$, because the proof relies on the recursion without projections, but for $\alpha = 1$, the rate of $O(n^{-1})$ (up to a logarithmic term) can be achieved in the more general framework of online learning, where averaging is key to deriving bounds for stochastic approximation from regret bounds. Moreover, bonds are obtained in high probability rather than simply in quadratic mean (see, e.g., [11, 12, 13, 14, 15]).

4 Non-strongly convex objectives

In this section, we do not assume that the function f is strongly convex, but we replace **(H3)** by:

(**H9**) The function f attains its global minimum at a certain $\theta^* \in \mathcal{H}$ (which may not be unique).

In the machine learning scenario, this essentially implies that the best predictor is in the function class we consider.[1] In the following theorem, since θ^* is not unique, we only derive a bound on function values. Not assuming strong convexity is essential in practice to make sure that algorithms are robust and *adaptive* to the hardness of the learning or optimization problem (much like gradient descent is).

4.1 Stochastic gradient descent

The following theorem is shown in a similar way to Theorem 1; we first derive a deterministic recursion, which we analyze with novel tools compared to the non-stochastic case (see details in the appendix):

Theorem 4 (Stochastic gradient descent, no strong convexity) *Assume* (**H1**,**H2'**,**H4**,**H9**). *Then, if* $\gamma_n = Cn^{-\alpha}$, *for* $\alpha \in [1/2, 1]$, *we have:*

$$\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right] \leqslant \frac{1}{C}\Big(\delta_0 + \frac{\sigma^2}{L^2}\Big) \exp\left(4L^2C^2\varphi_{1-2\alpha}(n)\right) \frac{1 + 4L^{3/2}C^{3/2}}{\min\{\varphi_{1-\alpha}(n), \varphi_{\alpha/2}(n)\}}. \tag{10}$$

When $\alpha = 1/2$, the bound goes to zero only when $LC < 1/4$, at rates which can be arbitrarily slow. For $\alpha \in (1/2, 2/3)$, we get convergence at rate $O(n^{-\alpha/2})$, while for $\alpha \in (2/3, 1)$, we get a convergence rate of $O(n^{\alpha-1})$. For $\alpha = 1$, the upper bound is of order $O((\log n)^{-1})$, which may be very slow (but still convergent). The rate of convergence changes at $\alpha = 2/3$, where we get our best rate $O(n^{-1/3})$, which does not match the minimax rate of $O(n^{-1/2})$ for stochastic approximation in the non-strongly convex case [23]. These rates for stochastic gradient descent without strong convexity assumptions are new and we conjecture that they are asymptotically minimax optimal (for stochastic gradient descent, not for stochastic approximation). Nevertheless, the proof of this result falls out of the scope of this paper.

If we further assume that we have all gradients bounded by B (that is, we assume $D = \infty$ in (**H5**)), then, we have the following theorem, which allows $\alpha \in (1/3, 1/2)$ with rate $O(n^{-3\alpha/2+1/2})$:

[1] For least-squares regression with kernels, where $f_n(\theta) = \frac{1}{2}(y_n - \langle \theta, \Phi(x_n)\rangle)^2$, with $\Phi(x_n)$ being the feature map associated with a reproducing kernel Hilbert space $\mathcal{H}$ with universal kernel [26], then we need that $x \mapsto \mathbb{E}(Y|X = x)$ is a function within the RKHS. Taking care of situations where this is not true is clearly of importance but out of the scope of this paper.

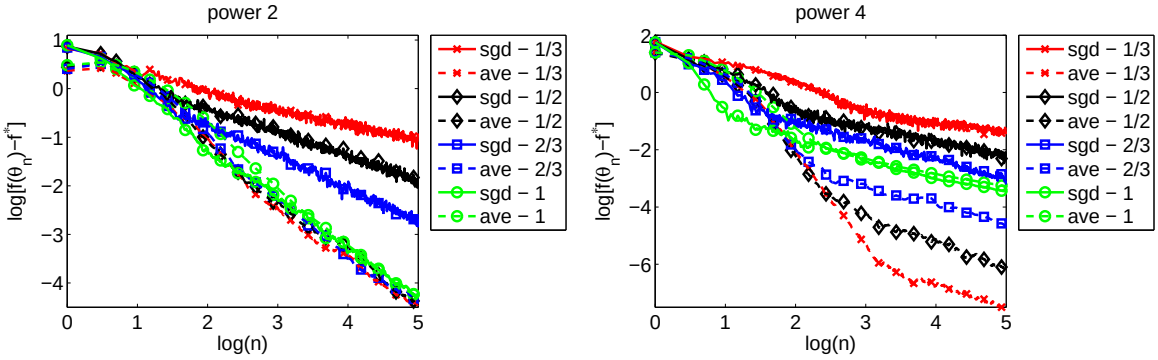

Figure 1: Robustness to lack of strong convexity for different learning rates and stochastic gradient (sgd) and Polyak-Ruppert averaging (ave). From left to right: $f(\theta) = |\theta|^2$ and $f(\theta) = |\theta|^4$, (between -1 and 1, affine outside of $[-1, 1]$, continuously differentiable). See text for details.

Theorem 5 (Stochastic gradient descent, no strong convexity, bounded gradients) *Assume* (**H1**, **H2'**, **H5**, **H9**). *Then, if* $\gamma_n = Cn^{-\alpha}$, *for* $\alpha \in [1/3, 1]$, *we have:*

$$\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right] \leqslant \begin{cases} \left(\delta_0 + B^2C^2\varphi_{1-2\alpha}(n)\right) \frac{1+4L^{1/2}C^{1/2}}{C\min\{\varphi_{1-\alpha}(n), \varphi_{\alpha/2}(n)\}}, & \text{if } \alpha \in [1/2, 1], \\ \frac{2}{C}(\delta_0 + B^2C^2)^{1/2} \frac{(1+4L^{1/2}BC^{3/2})}{(1-2\alpha)^{1/2}\varphi_{3\alpha/2-1/2}(n)}, & \text{if } \alpha \in [1/3, 1/2]. \end{cases} \tag{11}$$

4.2 Polyak-Ruppert averaging

Averaging in the context of non-strongly convex functions has been studied before, in particular in the optimization and machine learning literature, and the following theorems are similar in spirit to earlier work [7, 8, 13, 14, 15]:

Theorem 6 (averaging, no strong convexity) *Assume* (**H1**,**H2'**,**H4**,**H9**). *Then, if* $\gamma_n = Cn^{-\alpha}$, *for* $\alpha \in [1/2, 1]$, *we have*

$$\mathbb{E}\left[f(\bar{\theta}_n) - f(\theta^*)\right] \leqslant \frac{1}{C}\left(\delta_0 + \frac{\sigma^2}{L^2}\right) \frac{\exp\left(2L^2C^2\varphi_{1-2\alpha}(n)\right)}{n^{1-\alpha}} \left[1 + (2LC)^{1+\frac{1}{\alpha}}\right] + \frac{\sigma^2 C}{2n}\varphi_{1-\alpha}(n). \tag{12}$$

If $\alpha = 1/2$, then we only have convergence under $LC < 1/4$ (as in Theorem 4), with potentially slow rate, while for $\alpha > 1/2$, we have a rate of $O(n^{-\alpha})$, with otherwise similar behavior than for the strongly convex case with no bounded gradients. Here, averaging has allowed the rate to go from $O(\max\{n^{\alpha-1}, n^{-\alpha/2}\})$ to $O(n^{-\alpha})$.

Theorem 7 (averaging, no strong convexity, bounded gradients) *Assume* (**H1**,**H2'**,**H5**,**H9**). *If* $\gamma_n = Cn^{-\alpha}$, *for* $\alpha \in [0, 1]$, *we have*

$$\mathbb{E}\left[f(\bar{\theta}_n) - f(\theta^*)\right] \leqslant \frac{n^{\alpha-1}}{2C}(\delta_0 + C^2B^2\varphi_{1-2\alpha}(n)) + \frac{B^2}{2n}\varphi_{1-\alpha}(n). \tag{13}$$

With the bounded gradient assumption (and in fact without smoothness), we obtain the minimax asymptotic rate $O(n^{-1/2})$ up to logarithmic terms [23] for $\alpha = 1/2$, and, for $\alpha < 1/2$, the rate $O(n^{-\alpha})$ while for $\alpha > 1/2$, we get $O(n^{\alpha-1})$. Here, averaging has also allowed to increase the range of α which ensures convergence, to $\alpha \in (0, 1)$.

5 Experiments

In this section, we illustrate our theoretical results on synthetic and non-synthetic examples.

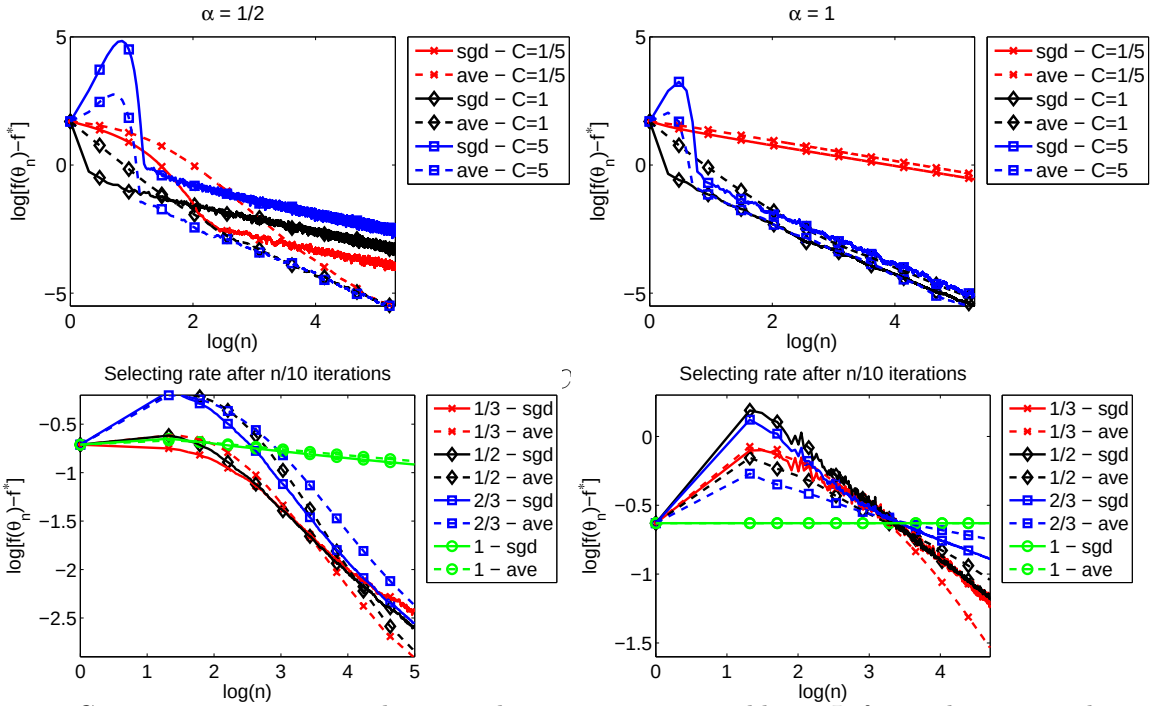

Fig
det

Figure 3: Comparison on non strongly convex logistic regression problems. Left: synthetic example, right: "alpha" dataset. See text for details. Best seen in color.

Robustness to lack of strong convexity. Define $f : \mathbb{R} \to \mathbb{R}$ as $|\theta|^q$ for $|\theta| \leqslant 1$ and extended into a continuously differentiable function, affine outside of $[-1, 1]$. For all $q > 1$, we have a convex function with Lipschitz-continuous gradient with constant $L = q(q-1)$. It is strongly convex around the origin for $q \in (1, 2]$, but its second derivative vanishes for $q > 2$. In Figure 1, we plot in log-log scale the average of $f(\theta_n) - f(\theta^*)$ over 100 replications of the stochastic approximation problem (with i.i.d. Gaussian noise of standard deviation 4 added to the gradient). For $q = 2$ (left plot), where we locally have a strongly convex case, all learning rates lead to good estimation with decay proportional to α (as shown in Theorem 1), while for the averaging case, all reach the exact same convergence rate (as shown in Theorem 3). However, for $q = 4$ where strong convexity does not hold (right plot), without averaging, $\alpha = 1$ is still fastest but becomes the slowest after averaging; on the contrary, illustrating Section 4, slower decays (such as $\alpha = 1/2$) leads to faster convergence when averaging is used. Note also the reduction in variability for the averaged iterations.

Robustness to wrong constants. We consider the function on the real line f, defined as $f(\theta) = \frac{1}{2}|\theta|^2$ and consider stochastic approximation with standard i.i.d. Gaussian noise on the gradients. In Figure 2, we plot the average performance over 100 replications, for various values of C and α. Note that for $\alpha = 1/2$ (left plot), the 3 curves for stochastic gradient descent end up being aligned and equally spaced, corroborating a rate proportional to C (see Theorem 1). Moreover, when averaging for $\alpha = 1/2$, the error ends up being independent of C and α (see Theorem 3). Finally, when C is too large, there is indeed an explosion (up to 10^5), hinting at the potential instability of having C too large. For $\alpha = 1$ (right plot), if C is too small, convergence is very slow (and not at the rate n^{-1}), as already observed by several authors (see, e.g., [8, 6]).

Medium-scale experiments with linear logistic regression. We consider two situations where $\mathcal{H} = \mathbb{R}^p$: (a) the "alpha" dataset from the Pascal large scale learning challenge (`http://largescale.ml.tu-berlin.de/`), for which $p = 500$ and $n = 50000$, and (b) a synthetic example where $p = 100$, $n = 100000$; we generate the input data i.i.d. from a multivariate Gaussian distribution with mean

zero and a covariance matrix sampled from a Wishart distribution with p degrees of freedom (thus with potentially bad condition number), and the output is obtained through a classification by a random hyperplane. For different values of α, we choose C in an adaptive way where we consider the lowest test error after $n/10$ iterations, and report results in Figure 3. In experiments reported in the appendix, we also consider a fixed rate equal to $1/L$ suggested by our analysis to avoid large constants, for which the convergence speed is very slow, suggesting that our global bounds involving the Lipschitz constants may be locally far too pessimistic and that designing a truly adaptive sequence (γ_n) instead of a fixed one is a fruitful avenue for future research.

6 Conclusion

In this paper, we have provided a non-asymptotic analysis of stochastic gradient, as well as its averaged version, for various learning rate sequences of the form $\gamma_n = Cn^{-\alpha}$ (see summary of results in Table 1). Following earlier work from the optimization, machine learning and stochastic approximation literatures, our analysis highlights that $\alpha = 1$ is not robust to the choice of C and to the actual difficulty of the problem (strongly convex or not). However, when using averaging with $\alpha \in (1/2, 1)$, we get, both in strongly convex and non-strongly convex situation, close to optimal rates of convergence. Our work can be extended in several ways: first, we have focused on results in quadratic mean and we expect that some of our results can be extended to results in high probability (in the line of [13, 3]). Second, we have focused on differentiable objectives, but the extension to objective functions with a differentiable stochastic part and a non-differentiable deterministic (in the line of [14]) would allow an extension to sparse methods. Finally, there are potentially further links between machine learning and the existing literature on stochastic approximation (see, e.g., [2]), which are worth pursuing.

α	SGD μ, L	SGD μ, B	SGD L	SGD L, B	Aver. μ, L	Aver. L	Aver. B
$(0, 1/3)$	α	α	$\times$	$\times$	2α	$\times$	α
$(1/3, 1/2)$	α	α	$\times$	$(3\alpha - 1)/2$	2α	$\times$	α
$(1/2, 2/3)$	α	α	$\alpha/2$	$\alpha/2$	1	$1 - \alpha$	$1 - \alpha$
$(2/3, 1)$	α	α	$1 - \alpha$	$1 - \alpha$	1	$1 - \alpha$	$1 - \alpha$

Table 1: Summary of results: For stochastic gradient descent (SGD) or Polyak-Ruppert averaging (Aver.), we provide their rates of convergence of the form $n^{-\beta}$ corresponding to learning rate sequences $\gamma_n = Cn^{-\alpha}$, where β is shown as a function of α. For each method, we list the main assumptions (μ: strong convexity, L: bounded Hessian, B: bounded gradients). For all columns but the second, we consider the convergence rates of function values.

Acknowledgement

This paper was partially supported by the European Research Council (SIERRA Project). We thank Mark Schmidt and Nicolas Le Roux for helpful discussions related to stochastic gradient descent algorithms.

In this appendix, we provide detailed proofs to all the results presented in the main paper. We also provide an additional simulation experiment comparing fixed learning rate sequences of the form $\gamma_n = Cn^{-\alpha}$, and adaptive ways of setting C.

A Proof of Theorem 1

Proof

Sketch. Under the stated assumptions, it is easily shown that (δ_n) satisfies the following recursion:

$$\delta_n \leqslant (1 - 2\mu\gamma_n + 2L^2\gamma_n^2)\delta_{n-1} + 2\sigma^2\gamma_n^2. \tag{14}$$

Note that it also appears in [3, Eq. (2)] under different assumptions. Using this recursion, we then derive bounds using some classical techniques from stochastic approximation [2], but in a non-asymptotic way, by deriving explicit upper-bounds.

Derivation of recursion. Using the Lipschitz continuity of f_n' (assumption **(H2)**), together with assumption **(H4)** and the fact that θ_{n-1} is $\mathcal{F}_{n-1}$-measurable, we obtain

$$\begin{aligned}\mathbb{E}(\|f_n'(\theta_{n-1})\|^2|\mathcal{F}_{n-1}) &\leqslant 2\mathbb{E}(\|f_n'(\theta^*) - f_n'(\theta_{n-1})\|^2|\mathcal{F}_{n-1}) + 2\mathbb{E}(\|f_n'(\theta^*)\|^2|\mathcal{F}_{n-1}) \\ &\leqslant 2L^2\|\theta_{n-1} - \theta^*\|^2 + 2\sigma^2.\end{aligned}$$

On the other hand, the recursion in Eq. (1) implies:

$$\|\theta_n - \theta^*\|^2 = \|\theta_{n-1} - \theta^*\|^2 - 2\gamma_n\langle\theta_{n-1} - \theta^*, f_n'(\theta_{n-1})\rangle + \gamma_n^2\|f_n'(\theta_{n-1})\|^2. \tag{15}$$

By computing the conditional expectation of the two sides of the previous equation, and using assumption (**H1**) together with the strong convexity assumption (assumption **(H3)**), in particular through [20, Eq. (2.1.17)] (i.e., for all θ_1, θ_2, $\langle f'(\theta_1) - f'(\theta_2), \theta_1 - \theta_2\rangle \geqslant \mu\|\theta_1 - \theta_2\|^2$)

$$\begin{aligned}\mathbb{E}(\|\theta_n - \theta^*\|^2|\mathcal{F}_{n-1}) &= \|\theta_{n-1} - \theta^*\|^2 - 2\gamma_n\langle\theta_{n-1} - \theta^*, f'(\theta_{n-1})\rangle + \gamma_n^2\mathbb{E}(\|f_n'(\theta_{n-1})\|^2|\mathcal{F}_{n-1}) \\ &\leqslant \|\theta_{n-1} - \theta^*\|^2 - 2\gamma_n\mu\|\theta_{n-1} - \theta^*\|^2 + \gamma_n^2\mathbb{E}(\|f_n'(\theta_{n-1})\|^2|\mathcal{F}_{n-1}) \\ &\leqslant (1 - 2\mu\gamma_n + 2L^2\gamma_n^2) \times \|\theta_{n-1} - \theta^*\|^2 + 2\sigma^2\gamma_n^2.\end{aligned}$$

Thus, setting $\delta_n = \mathbb{E}\|\theta_n - \theta^*\|^2$ and taking the expectation of the both side of the previous inequality yields to the following deterministic recursion:

$$\delta_n \leqslant (1 - 2\mu\gamma_n + 2L^2\gamma_n^2)\delta_{n-1} + 2\sigma^2\gamma_n^2. \tag{16}$$

General study of recursion for any decreasing sequence $(\gamma_n)_{n\in\mathbb{N}}$. Since $\mu \leqslant L$, we have $2\mu\gamma_n - 2L^2\gamma_n^2 \leqslant 2L\gamma_n - 2L^2\gamma_n^2 = 2L\gamma_n(1 - L\gamma_n) \leqslant 1/2$, thus the multiplicative factor in front of δ_{n-1} is always positive (in fact greater than $1/2$). Moreover, since all terms are positive, we have by applying the recursion n times:

$$\delta_n \leqslant \prod_{k=1}^{n}(1 - 2\mu\gamma_k + 2L^2\gamma_k^2)\delta_0 + 2\sigma^2\sum_{k=1}^{n}\prod_{i=k+1}^{n}(1 - 2\mu\gamma_i + 2L^2\gamma_i^2)\gamma_k^2. \tag{17}$$

The quadratic risk is therefore a sum of two terms: a *transient term*, depending only on the initial condition δ_0, of the form $A_{1,n}\delta_0$, and *stationary term* depending only on the noise variance, of the form $2\sigma^2 A_{2,n}$.

To deal with the transient term, we use the simple bound $1 + t \leqslant \exp(t)$ for any $t \in \mathbb{R}$, to obtain

$$A_{1,n} \stackrel{\text{def}}{=} \prod_{k=1}^{n}(1 - 2\mu\gamma_k + 2L^2\gamma_k^2) \leqslant \exp\left(-2\mu\sum_{k=1}^{n}\gamma_k\right)\exp\left(2L^2\sum_{k=1}^{n}\gamma_k^2\right). \tag{18}$$

For the stationary term, we consider two cases, depending whether γ_n is smaller or larger than $\frac{\mu}{2L^2}$, leading to two regimes, the first one where in fact the error δ_n may grow, and a second one where it goes back to zero.

Indeed, when $\gamma_n \leqslant \frac{\mu}{2L^2}$, then $1 - 2\mu\gamma_n + 2L^2\gamma_n^2 \leqslant 1 - \mu\gamma_n$ and in all cases, $1 - 2\mu\gamma_n + 2L^2\gamma_n^2 \leqslant 1 + 2L^2\gamma_n^2$. We denote by $n_0 = \inf\left\{n \in \mathbb{N}, \gamma_n \leqslant \frac{\mu}{2L^2}\right\}$. We then have:

$$\begin{aligned}
A_{2,n} &\overset{\text{def}}{=} \sum_{k=1}^{n} \prod_{i=k+1}^{n} (1 - 2\mu\gamma_i + 2L^2\gamma_i^2)\gamma_k^2 \\
&= \sum_{k=n_0+1}^{n} \prod_{i=k+1}^{n} (1 - 2\mu\gamma_i + 2L^2\gamma_i^2)\gamma_k^2 \\
&\quad + \sum_{k=1}^{n_0} \prod_{i=k+1}^{n_0} (1 - 2\mu\gamma_i + 2L^2\gamma_i^2)\gamma_k^2 \prod_{i=n_0+1}^{n} (1 - 2\mu\gamma_i + 2L^2\gamma_i^2) \qquad (19) \\
&\leqslant \sum_{k=1}^{n} \prod_{i=k+1}^{n} (1 - \mu\gamma_i)\gamma_k^2 + \left[\sum_{k=1}^{n_0} \prod_{i=k+1}^{n_0} (1 + 2L^2\gamma_i^2)\gamma_k^2\right] \prod_{i=n_0+1}^{n} (1 - \mu\gamma_i). \qquad (20)
\end{aligned}$$

We have:

$$\begin{aligned}
\sum_{k=1}^{n_0} \prod_{i=k+1}^{n_0} (1 + 2L^2\gamma_i^2)\gamma_k^2 &= \frac{1}{2L^2} \sum_{k=1}^{n_0} \left[\prod_{i=k}^{n_0} (1 + 2L^2\gamma_i^2) - \prod_{i=k+1}^{n_0} (1 + 2L^2\gamma_i^2)\right] \\
&\leqslant \frac{1}{2L^2} \prod_{i=1}^{n_0} (1 + 2L^2\gamma_i^2) \leqslant \frac{1}{2L^2} \exp\left(2L^2 \sum_{k=1}^{n_0} \gamma_k^2\right). \qquad (21)
\end{aligned}$$

Moreover, because for $n \leqslant n_0$, $\gamma_n \geqslant \frac{\mu}{2L^2}$, we obtain:

$$\begin{aligned}
\prod_{i=n_0+1}^{n} (1 - \mu\gamma_i) &= \exp\left(-\mu \sum_{i=1}^{n} \gamma_i\right) \exp\left(\mu \sum_{i=1}^{n_0} \gamma_i\right) \\
&\leqslant \exp\left(-\mu \sum_{i=1}^{n} \gamma_i\right) \exp\left(2L^2 \sum_{i=1}^{n_0} \gamma_i^2\right). \qquad (22)
\end{aligned}$$

Moreover, since (γ_n) is decreasing, for any $m \in \{1, \ldots, n\}$, we may split the following sum as follows:

$$\begin{aligned}
\sum_{k=1}^{n} \prod_{i=k+1}^{n} (1 - \mu\gamma_i)\gamma_k^2 &= \sum_{k=1}^{m} \prod_{i=k+1}^{n} (1 - \mu\gamma_i)\gamma_k^2 + \sum_{k=m+1}^{n} \prod_{i=k+1}^{n} (1 - \mu\gamma_i)\gamma_k^2 \\
&\leqslant \prod_{i=m+1}^{n} (1 - \mu\gamma_i) \sum_{k=1}^{m} \gamma_k^2 + \gamma_m \sum_{k=m+1}^{n} \prod_{i=k+1}^{n} (1 - \mu\gamma_i)\gamma_k \\
&\leqslant \exp\left(-\mu \sum_{i=m+1}^{n} \gamma_i\right) \sum_{k=1}^{m} \gamma_k^2 + \frac{\gamma_m}{\mu} \sum_{k=m+1}^{n} \left[\prod_{i=k+1}^{n} (1 - \mu\gamma_i) - \prod_{i=k}^{n} (1 - \mu\gamma_i)\right] \\
&\leqslant \exp\left(-\mu \sum_{i=m+1}^{n} \gamma_i\right) \sum_{k=1}^{m} \gamma_k^2 + \frac{\gamma_m}{\mu} \left[1 - \prod_{i=m+1}^{n} (1 - \mu\gamma_i)\right] \\
&\leqslant \exp\left(-\mu \sum_{i=m+1}^{n} \gamma_i\right) \sum_{k=1}^{n} \gamma_k^2 + \frac{\gamma_m}{\mu}.
\end{aligned}$$

This leads to the following bound on $A_{2,n}$:

$$A_{2,n} \leqslant \exp\left(-\mu \sum_{i=m+1}^{n} \gamma_i\right) \sum_{k=1}^{n} \gamma_k^2 + \frac{\gamma_m}{\mu} + \frac{1}{2L^2} \exp\left(4L^2 \sum_{k=1}^{n_0} \gamma_k^2\right) \exp\left(-\mu \sum_{i=1}^{n} \gamma_i\right), \tag{23}$$

with an alternative equation (with no split using m, which will be used for $\gamma_n = C/n$):

$$A_{2,n} \leqslant \sum_{k=1}^{n} \prod_{i=k+1}^{n} (1 - \mu\gamma_i)\gamma_k^2 + \frac{1}{2L^2} \exp\left(4L^2 \sum_{k=1}^{n_0} \gamma_k^2\right) \exp\left(-\mu \sum_{i=1}^{n} \gamma_i\right). \tag{24}$$

Note that in order for the previous inequalities to be valid for all n, we may simply replace n_0 by n (which we do later).

Study of recursion for $\gamma_n = Cn^{-\alpha}$, $\alpha \in (0, 1]$. We use the following inequalities (which are standardly obtained by bounding sums with integrals, and which could clearly be improved to obtain sharper bounds):

$$\forall \beta \geqslant 1, \quad \varphi_\beta(n) - \varphi_\beta(m) \leqslant \sum_{k=m+1}^{n} k^{\beta-1} \leqslant 2[\varphi_\beta(n) - \varphi_\beta(m)],$$

$$\forall \beta \leqslant 1, \quad \frac{1}{2}[\varphi_\beta(n) - \varphi_\beta(m)] \leqslant \sum_{k=m+1}^{n} k^{\beta-1} \leqslant \varphi_\beta(n) - \varphi_\beta(m).$$

We take bounds on the two terms $A_{1,n}$ and $A_{2,n}$. From Eq. (18), we get for all $\alpha \in (0, 1]$,

$$A_{1,n} \leqslant \exp[-\mu C \varphi_{1-\alpha}(n)] \ \exp[2L^2 C^2 \varphi_{1-2\alpha}(n)].$$

For any $\alpha \in (0, 1)$, Eq. (23) leads to

$$\begin{aligned} A_{2,n} &\leqslant \frac{2C}{\mu n^\alpha} + C^2 \varphi_{1-2\alpha}(n) \exp\left(-\frac{\mu C}{2}[\varphi_{1-\alpha}(n) - \varphi_{1-\alpha}(n/2)]\right) \\ &\qquad + \frac{1}{2L^2} \exp\left(4C^2L^2\varphi_{1-2\alpha}(n_0)\right) \exp\left(-\frac{\mu C}{2}\varphi_{1-\alpha}(n)\right) \\ &\leqslant \frac{2C}{\mu n^\alpha} + C^2 \varphi_{1-2\alpha}(n) \exp\left(-\frac{\mu C}{4} n^{1-\alpha}\right) \\ &\qquad + \frac{1}{2L^2} \exp\left(4C^2L^2\varphi_{1-2\alpha}(n_0)\right) \exp\left(-\frac{\mu C}{2}\varphi_{1-\alpha}(n)\right), \end{aligned}$$

where we have used the fact that for all $\beta \in [0, 1)$, $\varphi_\beta(x) - \varphi_\beta(x/2) \geqslant \frac{1}{2}x^\beta$. We can replace n_0 by n, thus leading to

$$\begin{aligned} \delta_n \leqslant \exp\left(4L^2C^2\varphi_{1-2\alpha}(n)\right) \exp\left(-\frac{\mu C}{2}\varphi_{1-\alpha}(n)\right) \left(\delta_0 + \frac{\sigma^2}{L^2}\right) \\ + \frac{4C\sigma^2}{\mu n^\alpha} + 2C^2\sigma^2 \varphi_{1-2\alpha}(n) \exp\left(-\frac{\mu C}{4} n^{1-\alpha}\right). \end{aligned}$$

We can now use the inequality $\varphi_{1-\alpha}(n) \geqslant n^{1-\alpha}$ to combine the first and third term into

$$2 \exp\left(4L^2C^2\varphi_{1-2\alpha}(n)\right) \exp\left(-\frac{\mu C}{4} n^{1-\alpha}\right) \left(\delta_0 + \frac{\sigma^2}{L^2}\right),$$

leading to Eq. (5).

For $\alpha = 1$, we get from Eq. (24),

$$\begin{aligned} A_{2,n} &\leqslant 2\sigma^2 \sum_{k=1}^{n} \exp\left(-\frac{\mu C}{2}\ln(n) + \frac{\mu C}{2}\ln(k)\right)\frac{C^2}{k^2} \\ &\qquad + \frac{1}{2L^2}\exp\left(4C^2L^2\varphi_{1-2\alpha}(n_0)\right)\exp\left(-\frac{\mu C}{2}\varphi_{1-\alpha}(n)\right). \\ &\leqslant 2\sigma^2 C^2 n^{-\mu C/2}\varphi_{\mu C/2-1}(n) + \frac{1}{2L^2}\exp\left(4C^2L^2\varphi_{1-2\alpha}(n_0)\right)\exp\left(-\frac{\mu C}{2}\varphi_{1-\alpha}(n)\right). \end{aligned}$$

Bound on function values. Using the Cauchy-Schwarz inequality, we get:

$$\begin{aligned} \mathbb{E}[f_n(\theta_{n-1}) - f_n(\theta^*)|\mathcal{F}_{n-1}] &= \int_0^1 \mathbb{E}\big[\langle f_n'(t\theta_{n-1} + (1-t)\theta^*) - f_n'(\theta^*), \theta_{n-1} - \theta^*\rangle|\mathcal{F}_{n-1}\big]\mathrm{d}t \\ &\leqslant \int_0^1 \Big[\mathbb{E}\big(\|f_n'(t\theta_{n-1} + (1-t)\theta^*) - f_n'(\theta^*)\|^2|\mathcal{F}_{n-1}\big)\Big]^{1/2}\|\theta_{n-1} - \theta^*\|\mathrm{d}t \\ &\leqslant L\|\theta_{n-1} - \theta^*\|^2 \int_0^1 t\mathrm{d}t = \frac{L}{2}\|\theta_{n-1} - \theta^*\|^2. \end{aligned}$$

This implies that, for all $n \in \mathbb{N}$, $\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right] \leqslant \frac{L}{2}\delta_n$.

■

B Proof of Theorem 2

Proof We follow the same proof technique than for Theorem 1. If we assume that all gradients are upperbounded, i.e., for all $n \in \mathbb{N}$ and all $\theta \in \mathcal{H}$, $\|f_n'(\theta)\| \leqslant B$, then the recursion in Eq. (6) becomes (using that orthogonal projectors are contractive and that $\|\theta^*\| \leqslant D$):

$$\delta_n \leqslant (1 - 2\mu\gamma_n)\delta_{n-1} + B^2\gamma_n^2 \leqslant \exp(-2\mu\gamma_n)\delta_{n-1} + B^2\gamma_n^2. \tag{25}$$

By applying Eq. (25) n times, we get

$$\delta_n \leqslant \delta_0 \prod_{k=1}^{n}\exp(-2\mu\gamma_k) + B^2\sum_{k=1}^{n}\gamma_k^2\prod_{i=k+1}^{n}\exp(-2\mu\gamma_i).$$

Case $\gamma_n = Cn^{-\alpha}$. For $\alpha < 1$, the first term leads to a bound,

$$\delta_0\exp[-\mu C\varphi_{1-\alpha}(n)]$$

while the second term leads to, following the same reasoning than for obtaining Eq. (23) in the proof of Theorem 1:

$$\frac{2B^2C^2}{\mu n^\alpha} + B^2C^2\varphi_{1-2\alpha}(n)\exp\left(-\frac{\mu C}{2}n^{1-\alpha}\right),$$

leading to an overall bound

$$\delta_0\exp[-\mu C\varphi_{1-\alpha}(n)] + \frac{2B^2C^2}{\mu' n^\alpha} + B^2C^2\varphi_{1-2\alpha}(n)\exp(-\mu C n^{1-\alpha}/2).$$

In order to get Eq. (8), we combine the first and third terms, by noticing that $\varphi_{1-\alpha}(n) \geqslant n^{1-\alpha}$.

For $\alpha = 1$, we get the bound (still following the same reasoning than for Theorem 1):

$$\delta_0 \exp[-\mu C \varphi_{1-\alpha}(n)] + 2B^2C^2 n^{-\mu C} \varphi_{\mu C - 1}(n).$$

■

C Proof of Theorem 3

Proof Following [4], we start from Eq. (1) and write it as $f'_n(\theta_{n-1}) = \frac{1}{\gamma_n}(\theta_{n-1} - \theta_n)$, and notice that:

(a) $f'_n(\theta_{n-1}) \approx f'_n(\theta^*) + f''(\theta^*)(\theta_{n-1} - \theta^*)$,

(b) $f'_n(\theta^*)$ has expectation zero and essentially behaves like an i.i.d. sequence,

(c) $\frac{1}{n}\sum_{k=1}^n \frac{1}{\gamma_k}(\theta_{k-1} - \theta_k)$ turns out to be negligible owing to a summation by parts and to the bound obtained in Theorem 1.

Thus $\bar{\theta}_n - \theta^*$ then behaves like $-\frac{1}{n}\sum_{k=1}^n f''(\theta^*)^{-1} f'_k(\theta^*)$, which leads to the bound in $O(n^{-1/2})$.

More precisely, we have:

$$\begin{aligned} f''(\theta^*)(\theta_{n-1} - \theta^*) &= f''_n(\theta^*)(\theta_{n-1} - \theta^*) + [f''(\theta^*) - f''_n(\theta^*)](\theta_{n-1} - \theta^*) \\ &= f'_n(\theta_{n-1}) - f'_n(\theta^*) + [f''_n(\theta^*)(\theta_{n-1} - \theta^*) - f'_n(\theta_{n-1}) + f'_n(\theta^*)] \\ &\quad + [f''(\theta^*) - f''_n(\theta^*)](\theta_{n-1} - \theta^*). \end{aligned}$$

Note that $f''(\theta^*)$ is invertible (with lowest eigenvalue greater than μ). We treat all terms separately:

– Note that we have for all $n \geqslant 1$, from Eq. (1):

$$f'_n(\theta_{n-1}) = \frac{1}{\gamma_n}(\theta_{n-1} - \theta_n).$$

We have, summing by parts,

$$\frac{1}{n}\sum_{k=1}^n \frac{1}{\gamma_k}(\theta_{k-1} - \theta_k) = \frac{1}{n}\sum_{k=1}^{n-1}(\theta_k - \theta^*)(\gamma_{k+1}^{-1} - \gamma_k^{-1}) - \frac{1}{n}(\theta_n - \theta^*)\gamma_n^{-1} + \frac{1}{n}(\theta_0 - \theta^*)\gamma_1^{-1},$$

leading to

$$\Big\|\frac{1}{n}\sum_{k=1}^n \frac{1}{\gamma_k}(\theta_{k-1} - \theta_k)\Big\| \leqslant \frac{1}{n}\sum_{k=1}^{n-1}\|\theta_k - \theta^*\| \cdot |\gamma_{k+1}^{-1} - \gamma_k^{-1}| + \frac{1}{n}\|\theta_n - \theta^*\|\gamma_n^{-1} + \frac{1}{n}\|\theta_0 - \theta^*\|\gamma_1^{-1}.$$

– Since $(f'_n(\theta^*))$ is a square-integrable martingale increment sequence in $\mathcal{H}$, we get

$$\mathbb{E}\Big\|f''(\theta^*)^{-1}\frac{1}{n}\sum_{k=0}^{n-1} f'_k(\theta^*)\Big\|^2 \leqslant \frac{1}{n}\operatorname{tr}\left[f''(\theta^*)^{-1}\Sigma f''(\theta^*)^{-1}\right].$$

– Because of assumption **(H5)**, we obtain

$$\|f_n''(\theta^*)(\theta_{n-1}-\theta^*) - f_n'(\theta_{n-1}) + f_n'(\theta^*)\| \leqslant \frac{M}{2}\|\theta_{n-1}-\theta^*\|^2.$$

– Similarly $[f''(\theta^*)-f_n''(\theta^*)](\theta_{n-1}-\theta^*)$ is also a martingale in $\mathcal{H}$, whose conditional second order moment is upper bounded by $4L^2\|\theta_{n-1}-\theta^*\|^2$. Thus

$$\mathbb{E}\left(\left\|\frac{1}{n}\sum_{k=1}^{n}[f''(\theta^*)-f_k''(\theta^*)](\theta_{k-1}-\theta^*)\right\|^2\right) \leqslant \frac{4L^2}{n^2}\sum_{k=0}^{n-1}\delta_k.$$

We finally obtain by combining all factors using Minkowski's inequality [17]:

$$\begin{aligned}\left(\mathbb{E}\|(\bar{\theta}_n-\theta^*)\|^2\right)^{1/2} &\leqslant \frac{(\operatorname{tr} f''(\theta^*)^{-1}\Sigma f''(\theta^*)^{-1})^{1/2}}{\sqrt{n}} + \frac{1}{n\gamma_n\mu^{1/2}}\delta_n^{1/2} + \frac{1}{n\gamma_1\mu^{1/2}}\delta_0^{1/2} + \frac{2L}{n\mu^{1/2}}\left(\sum_{k=0}^{n-1}\delta_k\right)^{1/2} \\ &\quad + \frac{1}{n\mu^{1/2}}\sum_{k=1}^{n-1}\delta_k^{1/2}|\gamma_{k+1}^{-1}-\gamma_k^{-1}| + \frac{M}{2n\mu^{1/2}}\sum_{k=1}^{n}(\mathbb{E}\|\theta_k-\theta^*\|^4)^{1/2}, \qquad (26)\end{aligned}$$

which can further simplify as

$$\begin{aligned}\left(\mathbb{E}\|(\bar{\theta}_n-\theta^*)\|^2\right)^{1/2} &\leqslant \frac{(\operatorname{tr} f''(\theta^*)^{-1}\Sigma f''(\theta^*)^{-1})^{1/2}}{\sqrt{n}} + \frac{1}{n\gamma_n\mu^{1/2}}\delta_n^{1/2} \\ &\quad + \frac{1}{n\mu^{1/2}}\left(\frac{1}{\gamma_1}+2L\right)\delta_0^{1/2} + \frac{2L}{n\mu^{1/2}}\left(\sum_{k=1}^{n}\delta_k\right)^{1/2} \qquad (27) \\ &\quad + \frac{1}{n\mu^{1/2}}\sum_{k=1}^{n-1}\delta_k^{1/2}|\gamma_{k+1}^{-1}-\gamma_k^{-1}| + \frac{M}{2n\mu^{1/2}}\sum_{k=1}^{n}(\mathbb{E}\|\theta_k-\theta^*\|^4)^{1/2} \qquad (28)\end{aligned}$$

In order to further bound the error, we can first re-use results from Theorem 1, but we also need to derive a bound on the fourth-order moment $\mathbb{E}\|\theta_k-\theta^*\|^4$.

Fourth-order moment. We now derive a recursion for the fourth-order moment, following the same arguments than for the second-order moment:

$$\begin{aligned}\|\theta_n-\theta^*\|^4 &= \left(\|\theta_{n-1}-\theta^*-\gamma_n f_n'(\theta_{n-1})\|^2\right)^2 \\ &= \left(\|\theta_{n-1}-\theta^*\|^2 + \gamma_n^2\|f_n'(\theta_{n-1})\|^2 - 2\gamma_n\langle\theta_{n-1}-\theta^*, f_n'(\theta_{n-1})\rangle\right)^2 \\ &= \|\theta_{n-1}-\theta^*\|^4 + 4\gamma_n^2\langle\theta_{n-1}-\theta^*, f_n'(\theta_{n-1})\rangle^2 + \gamma_n^4\|f_n'(\theta_{n-1})\|^4 \\ &\quad -4\gamma_n\|\theta_{n-1}-\theta^*\|^2\langle\theta_{n-1}-\theta^*, f_n'(\theta_{n-1})\rangle + 2\gamma_n^2\|\theta_{n-1}-\theta^*\|^2\|f_n'(\theta_{n-1})\|^2 \\ &\quad -4\gamma_n^3\langle\theta_{n-1}-\theta^*, f_n'(\theta_{n-1})\rangle\|f_n'(\theta_{n-1})\|^2.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}\left[\|\theta_n-\theta^*\|^4\middle|\mathcal{F}_{n-1}\right] &\leqslant \|\theta_{n-1}-\theta^*\|^4 + 6\gamma_n^2\|\theta_{n-1}-\theta^*\|^2\mathbb{E}\left[\|f_n'(\theta_{n-1})\|^2\middle|\mathcal{F}_{n-1}\right] \qquad (29) \\ &\quad +\gamma_n^4\mathbb{E}\left[\|f_n'(\theta_{n-1})\|^4\middle|\mathcal{F}_{n-1}\right] \\ &\quad -4\gamma_n\|\theta_{n-1}-\theta^*\|^2\langle\theta_{n-1}-\theta^*, f'(\theta_{n-1})-f'(\theta^*)\rangle \\ &\quad +4\gamma_n^3\|\theta_{n-1}-\theta^*\|\mathbb{E}\left[\|f_n'(\theta_{n-1})\|^3\middle|\mathcal{F}_{n-1}\right].\end{aligned}$$

Since the function f is strongly convex,

$$\langle\theta_{n-1}-\theta^*, f'(\theta_{n-1})-f'(\theta^*)\rangle \geq \mu\|\theta_{n-1}-\theta^*\|^2. \qquad (30)$$

On the other hand, the decomposition

$$\|f_n'(\theta_{n-1})\| \leqslant \|f_n'(\theta_{n-1}) - f_n'(\theta^*) + f_n'(\theta^*)\| \leqslant L\|\theta_{n-1} - \theta^*\| + \|f_n'(\theta^*)\| \ ,$$

implies that for all $p \in \{1, \ldots, 4\}$

$$\mathbb{E}\left[\|f_n'(\theta_{n-1})\|^p | \mathcal{F}_{n-1}\right] \leqslant 2^{p-1}\left[L^p\|\theta_{n-1} - \theta^*\|^p + \tau^p\right], \tag{31}$$

where we have used that $\mathbb{E}(\|f_n'(\theta_*)\|^p|\mathcal{F}_{n-1}) \leqslant \tau^p$ (Assumption (**H6**)). Combining Eq. (29) with Eq. (30) and Eq. (31) yields

$$\begin{aligned}\mathbb{E}\left[\|\theta_n - \theta^*\|^4 \middle| \mathcal{F}_{n-1}\right] \leqslant \|\theta_{n-1} - \theta^*\|^4 \left[1 - 4\mu\gamma_n + 12\gamma_n^2L^2 + 16\gamma_n^3L^3 + 8\gamma_n^4L^4\right] \\ +12\|\theta_{n-1} - \theta^*\|^2\gamma_n^2\tau^2 + 16\|\theta_{n-1} - \theta^*\|\gamma_n^3\tau^3 + 8\gamma_n^4\tau^4 \ .\end{aligned} \tag{32}$$

Since

$$16\|\theta_{n-1} - \theta^*\|\gamma_n^3\tau^3 \leqslant 8\left[\gamma_n^2\tau^2\|\theta_{n-1} - \theta^*\|^2 + \gamma_n^4\tau^4\right] \ ,$$

Eq. (32) yields

$$\begin{aligned}\mathbb{E}\left[\|\theta_n - \theta^*\|^4 \middle| \mathcal{F}_{n-1}\right] \leqslant \|\theta_{n-1} - \theta^*\|^4 \left[1 - 4\mu\gamma_n + 12\gamma_n^2L^2 + 16\gamma_n^3L^3 + 8\gamma_n^4L^4\right] \\ +20\|\theta_{n-1} - \theta^*\|^2\gamma_n^2\tau^2 + 16\gamma_n^4\tau^4 \ .\end{aligned}$$

We can now combine this with Eq. (6) and replace σ^2 by $\tau^2 \geqslant \sigma^2$, i.e.,

$$\mathbb{E}\left[\|\theta_n - \theta^*\|^2 \middle| \mathcal{F}_{n-1}\right] \leqslant (1 - 2\mu\gamma_n + 2L^2\gamma_n^2)\|\theta_{n-1} - \theta^*\|^2 + 2\tau^2\gamma_n^2, \tag{33}$$

to derive a recursion for

$$u_n = \|\theta_n - \theta^*\|^4 + \frac{20\gamma_{n+1}\tau^2}{\mu}\|\theta_n - \theta^*\|^2.$$

Indeed, we have, using $\frac{1}{2}\gamma_n \leqslant \gamma_{n+1} \leqslant \gamma_n$:

$$\begin{aligned}\mathbb{E}(u_n|\mathcal{F}_{n-1}) &\leqslant \|\theta_{n-1} - \theta^*\|^4 \left[1 - 4\mu\gamma_n + 12\gamma_n^2L^2 + 16\gamma_n^3L^3 + 8\gamma_n^4L^4\right] \\ &\quad +\|\theta_{n-1} - \theta^*\|^2\Big[20\gamma_n^2\tau^2 + \frac{20\gamma_{n+1}\sigma^2}{\mu}(1 - 2\mu\gamma_n + 2L^2\gamma_n^2)\Big] + 16\gamma_n^4\tau^4 + \frac{40\gamma_{n+1}}{\mu}\tau^4\gamma_n^2 \\ &\leqslant u_{n-1}\left[1 - \mu\gamma_n + 12\gamma_n^2L^2 + 16\gamma_n^3L^3 + 8\gamma_n^4L^4\right] + 16\gamma_n^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_n^3 \\ &\quad +\|\theta_{n-1} - \theta^*\|^2\Big[20\gamma_n^2\tau^2 + \frac{20\gamma_n\tau^2}{\mu}(1 - 2\mu\gamma_n + 2L^2\gamma_n^2) \\ &\qquad\qquad +\frac{20\gamma_n\tau^2}{\mu}\left(-1 + \mu\gamma_n - 12\gamma_n^2L^2 - 16\gamma_n^3L^3 - 8\gamma_n^4L^4\right)\Big] \\ &\leqslant u_{n-1}\left[1 - \mu\gamma_n + 12\gamma_n^2L^2 + 16\gamma_n^3L^3 + 8\gamma_n^4L^4\right] + 16\gamma_n^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_n^3 \\ &\leqslant u_{n-1}\left[1 - \mu\gamma_n + 16\gamma_n^2L^2 + 24\gamma_n^4L^4\right] + 16\gamma_n^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_n^3.\end{aligned}$$

Following the proof of Theorem 1, we consider the following terms $A'_{1,n}$ and $A'_{2,n}$, where n_0 is the largest n such that $16\gamma_n^2 L^2 + 24\gamma_n^4 L^4 \geqslant \gamma_n \mu/2$:

$$\begin{aligned}
A'_{1,n} &\stackrel{\text{def}}{=} \exp\left(-\mu\sum_{k=1}^n \gamma_k + 16L^2\sum_{k=1}^n \gamma_k^2 + 24L^4\sum_{k=1}^n \gamma_k^4\right)\\
A'_{2,n} &\stackrel{\text{def}}{=} \sum_{k=n_0+1}^n \prod_{i=k+1}^n \exp(-\mu\gamma_i/2)\left(16\gamma_k^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_k^3\right)\\
&\quad + \prod_{k=n_0+1}^n \exp(-\mu\gamma_k/2)\sum_{k=1}^{n_0}\prod_{i=k+1}^{n_0}\exp(16\gamma_i^2L^2 + 24\gamma_i^4L^4)\left(16\gamma_k^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_k^3\right)
\end{aligned}$$

We have

$$\begin{aligned}
A'_{2,n} &\leqslant \sum_{k=1}^n \prod_{i=k+1}^n \exp(-\mu\gamma_i/2)\left(16\gamma_k^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_k^3\right)\\
&\quad + \prod_{k=1}^n \exp(-\mu\gamma_k/2)\prod_{k=1}^{n_0}\exp(\mu\gamma_k/2)\prod_{i=1}^{n_0}\exp(16\gamma_i^2L^2 + 24\gamma_i^4L^4)\sum_{k=1}^{n_0}\left(16\gamma_k^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_k^3\right)\\
&\leqslant \sum_{k=1}^n \prod_{i=k+1}^n \exp(-\mu\gamma_i/2)\left(16\gamma_k^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_k^3\right)\\
&\quad + \prod_{k=1}^{n_0}\exp(32\gamma_k^2L^2 + 48\gamma_k^4L^4)\prod_{k=1}^n \exp(-\mu\gamma_k/2)\sum_{k=1}^{n_0}\left(16\gamma_k^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_k^3\right)
\end{aligned}$$

Moreover, we further split $A'_{2,n}$ and get, for any $m \in (0, n)$:

$$\begin{aligned}
A'_{3,n} &\stackrel{\text{def}}{=} \sum_{k=1}^n \prod_{i=k+1}^n \exp(-\mu\gamma_i/2)\left(16\gamma_k^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_k^3\right)\\
&= \sum_{k=1}^m \prod_{i=k+1}^n \exp(-\mu\gamma_i/2)\left(16\gamma_k^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_k^3\right) + \sum_{k=m+1}^n \prod_{i=k+1}^n \exp(-\mu\gamma_i/2)\left(16\gamma_k^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_k^3\right)\\
&\leqslant \prod_{i=m+1}^n \exp(-\mu\gamma_i/2)\sum_{k=1}^m\left(16\gamma_k^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_k^3\right) + \sum_{k=m+1}^n \prod_{i=k+1}^n \exp(-\mu\gamma_i/2)\left(16\gamma_m^3\tau^4 + \frac{40}{\mu}\tau^4\gamma_m^2\right)\gamma_k\\
&\leqslant \prod_{i=m+1}^n \exp(-\mu\gamma_i/2)\sum_{k=1}^n\left(16\gamma_k^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_k^3\right) + \left(16\gamma_m^3\tau^4 + \frac{40}{\mu}\tau^4\gamma_m^2\right)\sum_{k=1}^n\prod_{i=k+1}^n \exp(-\mu\gamma_i/2)\gamma_k\\
&\leqslant \prod_{i=m+1}^n \exp(-\mu\gamma_i/2)\sum_{k=1}^n\left(16\gamma_k^4\tau^4 + \frac{40}{\mu}\tau^4\gamma_k^3\right) + \left(16\gamma_m^3\tau^4 + \frac{40}{\mu}\tau^4\gamma_m^2\right)\frac{1}{\mu}
\end{aligned}$$

This leads to, for $\gamma_n = Cn^{-\alpha}$ and $m = n/2$:

$$\begin{aligned}
u_n &\leqslant u_0 A'_{1,n} + A'_{2,n}\\
&\leqslant \exp\left(-\frac{\mu C}{2}\varphi_{1-\alpha}(n) + 16L^2C^2\varphi_{1-2\alpha}(n) + 24L^4C^4\varphi_{1-4\alpha}(n)\right)u_0\\
&\quad + \exp\left(-\frac{\mu C}{4}\varphi_{1-\alpha}(n) + 32L^2C^2\varphi_{1-2\alpha}(n) + 48L^4C^4\varphi_{1-4\alpha}(n)\right)\left(16\tau^4C^4\varphi_{1-4\alpha}(n) + \frac{40\tau^4C^3}{\mu}\varphi_{1-3\alpha}(n)\right)\\
&\quad + \frac{1}{\mu}\left(\frac{16\times 2^{3\alpha}C^3\tau^4}{n^{3\alpha}} + \frac{40\times 2^{4\alpha}\tau^4C^2}{\mu n^{2\alpha}}\right) + \left(16\tau^4C^4\varphi_{1-4\alpha}(n) + \frac{40\tau^4C^3}{\mu}\varphi_{1-3\alpha}(n)\right)\exp(-\frac{\mu C}{8}n^{1-\alpha}).
\end{aligned}$$

Since $\varphi_{1-4\alpha}(n) \leqslant 1$ and $\varphi_{1-3\alpha}(n) \leqslant 3$, we obtain:

$$\begin{aligned} u_n &\leqslant \exp\left(-\frac{\mu C}{2}\varphi_{1-\alpha}(n) + 16L^2C^2\varphi_{1-2\alpha}(n) + 24L^4C^4\right)u_0 \\ &\quad + \exp\left(-\frac{\mu C}{4}\varphi_{1-\alpha}(n) + 32L^2C^2\varphi_{1-2\alpha}(n) + 48L^4C^4\right)\left(16\tau^4C^4 + \frac{80\tau^4C^3}{\mu}\right) \\ &\quad + \frac{1}{\mu}\left(\frac{128C^3\tau^4}{n^{3\alpha}} + \frac{640\tau^4C^2}{\mu n^{2\alpha}}\right) + \left(16\tau^4C^4 + \frac{80\tau^4C^3}{\mu}\right)\exp(-\frac{\mu C}{8}n^{1-\alpha}). \end{aligned} \tag{34}$$

With further simplifications, we get:

$$\begin{aligned} u_n &\leqslant \exp\left(-\frac{\mu C}{8}n^{1-\alpha} + 32L^2C^2\varphi_{1-2\alpha}(n) + 48L^4C^4\right)\left(32\tau^4C^4 + \frac{160\tau^4C^3}{\mu} + \mathbb{E}\|\theta_0 - \theta^*\|^4 + \frac{20C\tau^2}{\mu}\delta_0\right) \\ &\quad + \frac{1}{\mu}\left(\frac{128C^3\tau^4}{n^{3\alpha}} + \frac{640\tau^4C^2}{\mu n^{2\alpha}}\right). \end{aligned} \tag{35}$$

Overall bound. For $\alpha < 1$, we now compute a bound on all terms from Eq. (26), using Theorem 1 and Eq. (34), with the notation

$$A = \sum_{k=1}^{n} \exp\left(-\frac{\mu C}{16}k^{1-\alpha} + 16L^2C^2\varphi_{1-2\alpha}(k)\right). \tag{36}$$

$$
\begin{aligned}
\frac{1}{n\gamma_n\mu^{1/2}}\delta_n^{1/2} &\leqslant \frac{1}{n^{1-\alpha}\mu^{1/2}C}\frac{2C^{1/2}\sigma}{\mu^{1/2}n^{\alpha/2}} \\
&\quad +2\frac{1}{n^{1-\alpha}\mu^{1/2}C}\exp\left(2L^2C^2\varphi_{1-2\alpha}(n)\right)\exp\left(-\frac{\mu C}{8}n^{1-\alpha}\right)\left(\delta_0+\frac{\sigma^2}{L^2}\right)^{1/2} \\
&\leqslant \frac{2\sigma}{\mu C^{1/2}n^{1-\alpha/2}}+\left(\delta_0+\frac{\sigma^2}{L^2}\right)^{1/2}\frac{2A}{n\mu^{1/2}C} \\
\frac{1}{n\mu^{1/2}}\left(\frac{1}{\gamma_1}+2L\right)\delta_0^{1/2} &\leqslant \frac{1}{n\mu^{1/2}}\left(\frac{1}{C}+2L\right)\delta_0^{1/2} \\
\frac{2L}{n\mu^{1/2}}\left(\sum_{k=1}^n\delta_k\right)^{1/2} &\leqslant \frac{2L}{n\mu^{1/2}}\Big[\frac{4C\sigma^2}{\mu}\varphi_{1-\alpha}(n) \\
&\quad +2\sum_{k=1}^n\exp\left(4L^2C^2\varphi_{1-2\alpha}(k)\right)\exp\left(-\frac{\mu C}{4}k^{1-\alpha}\right)\left(\delta_0+\frac{\sigma^2}{L^2}\right)\Big]^{1/2} \\
&\leqslant \frac{2L}{n\mu^{1/2}}\frac{2C^{1/2}\sigma}{\mu^{1/2}}\varphi_{1-\alpha}(n)^{1/2}+\frac{4L}{n\mu^{1/2}}\left(\delta_0+\frac{\sigma^2}{L^2}\right)^{1/2}A^{1/2} \\
\frac{1}{n\mu^{1/2}}\sum_{k=1}^{n-1}\delta_k^{1/2}|\gamma_{k+1}^{-1}-\gamma_k^{-1}| &\leqslant \frac{2\alpha}{Cn\mu^{1/2}}\sum_{k=1}^n\frac{2C^{1/2}\sigma}{\mu^{1/2}}k^{\alpha/2-1} \\
&\quad +\frac{4\alpha}{Cn\mu^{1/2}}\sum_{k=1}^n k^{\alpha-1}\exp\left(2L^2C^2\varphi_{1-2\alpha}(n)\right)\exp\left(-\frac{\mu C}{8}n^{1-\alpha}\right)\left(\delta_0+\frac{\sigma^2}{L^2}\right)^{1/2} \\
&\leqslant \frac{4\sigma\alpha}{C^{1/2}n\mu}\varphi_{\alpha/2}(n)+\frac{4\alpha}{Cn\mu^{1/2}}\left(\delta_0+\frac{\sigma^2}{L^2}\right)^{1/2}A \\
\frac{M}{2n\mu^{1/2}}\sum_{k=1}^n(\mathbb{E}\|\theta_k-\theta^*\|^4)^{1/2} &\leqslant \frac{M}{2n\mu^{1/2}}\sum_{k=1}^n\Big[\frac{1}{\mu^{1/2}}\frac{12C^{3/2}\tau^2}{k^{3\alpha/2}}+\frac{26\tau^2C}{\mu k^\alpha}\Big] \\
&\quad +\frac{M}{2n\mu^{1/2}}\sum_{k=1}^n\exp\left(-\frac{\mu C}{16}k^{1-\alpha}+16L^2C^2\varphi_{1-2\alpha}(k)+24L^4C^4\right) \\
&\qquad\qquad \times\left(32\tau^4C^4+\frac{160\tau^4C^3}{\mu}+\mathbb{E}\|\theta_0-\theta^*\|^4+\frac{20C\tau^2}{\mu}\delta_0\right)^{1/2} \\
&\leqslant \frac{MC\tau^2}{2n\mu}\Big[C^{1/2}\varphi_{1-3\alpha/2}(n)+\mu^{-1/2}\varphi_{1-\alpha}(n)\Big] \\
&\quad +\frac{M\sqrt{20}C^{1/2}\tau}{2n\mu}A\exp\left(24L^4C^4\right)\left(\delta_0+\frac{\mu\mathbb{E}\|\theta_0-\theta^*\|^4}{20C\tau^2}+2\tau^2C^3\mu+8\tau^2C^2\right)^{1/2}
\end{aligned}
$$

Behavior of the constant A. For all $\alpha\in(0,1)$, A is finite, while for $\alpha=1$, $A=O(n)$. Note that when CL is too large, A may be large as well.

Final bound. We get a bound of the form

$$\begin{aligned}\left(\mathbb{E}\|(\bar{\theta}_n-\theta^*)\|^2\right)^{1/2} \leqslant\ & \frac{(\operatorname{tr} f''(\theta^*)^{-1}\Sigma f''(\theta^*)^{-1})^{1/2}}{\sqrt{n}} \\ &+\frac{2\sigma}{\mu C^{1/2}n^{1-\alpha/2}}+\left(\delta_0+\frac{\sigma^2}{L^2}\right)^{1/2}\frac{2A}{n\mu^{1/2}C} \\ &+\frac{1}{n\mu^{1/2}}\left(\frac{1}{C}+2L\right)\delta_0^{1/2} \\ &+\frac{2L}{n\mu^{1/2}}\frac{2C^{1/2}\sigma}{\mu^{1/2}}\varphi_{1-\alpha}(n)^{1/2}+\frac{4L}{n\mu^{1/2}}\left(\delta_0+\frac{\sigma^2}{L^2}\right)^{1/2}A^{1/2} \\ &+\frac{4\sigma\alpha}{C^{1/2}n\mu}\varphi_{\alpha/2}(n)+\frac{4\alpha}{Cn\mu^{1/2}}\left(\delta_0+\frac{\sigma^2}{L^2}\right)^{1/2}A \\ &\frac{MC\tau^2}{2n\mu}\left[C^{1/2}\varphi_{1-3\alpha/2}(n)+\mu^{-1/2}\varphi_{1-\alpha}(n)\right] \\ &+\frac{M\sqrt{20}C^{1/2}\tau}{2n\mu}A\exp\left(24L^4C^4\right)\left(\delta_0+\frac{\mu\mathbb{E}\|\theta_0-\theta^*\|^4}{20C\tau^2}+2\tau^2C^3\mu+8\tau^2C^2\right)^{1/2}\end{aligned}$$

which we can simplify into

$$\begin{aligned}\left(\mathbb{E}\|(\bar{\theta}_n-\theta^*)\|^2\right)^{1/2} \leqslant\ & \frac{(\operatorname{tr} f''(\theta^*)^{-1}\Sigma f''(\theta^*)^{-1})^{1/2}}{\sqrt{n}} \\ &+\frac{6\sigma}{\mu C^{1/2}}\frac{1}{n^{1-\alpha/2}}+\frac{4LC^{1/2}}{\mu}\frac{\varphi_{1-\alpha}(n)^{1/2}}{n} \\ &+\frac{MC\tau^2}{2n\mu}\left[C^{1/2}\varphi_{1-3\alpha/2}(n)+\mu^{-1/2}\varphi_{1-\alpha}(n)\right] \\ &+\frac{8A}{n\mu^{1/2}}\left(\frac{1}{C}+L\right)\left(\delta_0+\frac{\sigma^2}{L^2}\right)^{1/2} \\ &+\frac{5MC^{1/2}\tau}{2n\mu}A\exp\left(24L^4C^4\right)\left(\delta_0+\frac{\mu\mathbb{E}\|\theta_0-\theta^*\|^4}{20C\tau^2}+2\tau^2C^3\mu+8\tau^2C^2\right)^{1/2}.\end{aligned}$$

We can further simplify by noticing that $\varphi_{1-3\alpha/2}(n) \leqslant \varphi_{1-\alpha}(n)$, and thus obtain the desired result. ■

D Proof of Theorem 4

Proof In this proof, we follow the proof technique of [20] for the deterministic case. Define

$$\Delta_n = \mathbb{E}\left[f(\theta_n) - f(\theta^*)\right]. \tag{37}$$

We also derive a deterministic bound, which we upper bound, but here on the function values Δ_n . We start by showing that the iterates remain bounded in quadratic mean.

Note that the bound that we derive depends on a particular choice of θ^* among all minimizers of f.

Bound on $\|\theta_n - \theta^*\|^2$. Following the same argument than for strongly convex functions to obtain Eq. (6) (but taking $\mu = 0$), we have, for $\delta_n = \mathbb{E}\|\theta_n - \theta^*\|^2$, using (**H1**), (**H2'**), (**H4**):

$$\delta_n \leqslant (1 + 2L^2\gamma_n^2)\delta_{n-1} + 2\sigma^2\gamma_n^2. \tag{38}$$

This implies, by applying the previous recursion n times, that

$$\begin{aligned}
\delta_n &\leqslant \prod_{k=1}^{n}(1+2L^2\gamma_k^2)\delta_0 + 2\sigma^2\sum_{k=1}^{n}\prod_{j=k+1}^{n}(1+2L^2\gamma_j^2)\gamma_k^2 \\
&= \prod_{k=1}^{n}(1+2L^2\gamma_k^2)\delta_0 + \frac{\sigma^2}{L^2}\sum_{k=1}^{n}\left[\prod_{j=k+1}^{n}(1+2L^2\gamma_j^2) - \prod_{j=k}^{n}(1+2L^2\gamma_j^2)\right] \\
&\leqslant \exp\left(2L^2\sum_{k=1}^{n}\gamma_k^2\right)\left(\delta_0 + \frac{\sigma^2}{L^2}\right) \stackrel{\text{def}}{=} D_n, \qquad (39)
\end{aligned}$$

which provides an upper-bound on the mean-square error δ_n. Moreover, we have, using the Lipschitz continuity of f, and taking conditional expectations:

$$\begin{aligned}
f(\theta_n) &\leqslant f(\theta_{n-1}) - \gamma_n\langle f'(\theta_{n-1}), f_n'(\theta_{n-1})\rangle + \frac{\gamma_n^2 L}{2}\|f_n'(\theta_{n-1})\|^2 \\
\mathbb{E}\left[f(\theta_n)\middle|\mathcal{F}_{n-1}\right] &\leqslant f(\theta_{n-1}) - \gamma_n\|f'(\theta_{n-1})\|^2 + \frac{\gamma_n^2 L}{2}\mathbb{E}\left[\|f_n'(\theta_{n-1})\|^2\middle|\mathcal{F}_{n-1}\right]. \qquad (40)
\end{aligned}$$

Using inequality (2.1.7) from [20] (owing to the Lipschitz continuity of f_n), we get:

$$\frac{1}{2L}\|f_n'(\theta_{n-1}) - f_n'(\theta^*)\|^2 \leqslant f_n(\theta_{n-1}) - f_n(\theta^*) - \langle f_n'(\theta^*), \theta_{n-1} - \theta^*\rangle.$$

By computing the conditional expectation, we therefore obtain, since θ_{n-1} is $\mathcal{F}_{n-1}$-measurable,

$$\frac{1}{2L}\mathbb{E}\left[\|f_n'(\theta_{n-1}) - f_n'(\theta^*)\|^2\middle|\mathcal{F}_{n-1}\right] \leqslant f(\theta_{n-1}) - f(\theta^*). \qquad (41)$$

Combining the two inequalities in Eq. (40) and Eq. (41) leads to

$$\begin{aligned}
&\mathbb{E}\left[f(\theta_n) - f(\theta^*)\middle|\mathcal{F}_{n-1}\right] \\
\leqslant\ & f(\theta_{n-1}) - f(\theta^*) - \gamma_n\|f'(\theta_{n-1})\|^2 + \frac{\gamma_n^2 L}{2}\mathbb{E}\left[\|f_n'(\theta_{n-1})\|^2\middle|\mathcal{F}_{n-1}\right] \\
\leqslant\ & f(\theta_{n-1}) - f(\theta^*) - \gamma_n\|f'(\theta_{n-1})\|^2 + \frac{\gamma_n^2 L}{2}\mathbb{E}\left[2\|f_n'(\theta^*)\|^2 + 2\|f_n'(\theta^*) - f_n'(\theta_{n-1})\|^2\middle|\mathcal{F}_{n-1}\right] \\
\leqslant\ & (1+2\gamma_n^2L^2)[f(\theta_{n-1}) - f(\theta^*)] - \gamma_n\|f'(\theta_{n-1})\|^2 + \gamma_n^2 L\sigma^2.
\end{aligned}$$

Taking the expectation on the both sides of the previous identity yields

$$\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right] \leqslant (1+2\gamma_n^2L^2)\mathbb{E}\left[f(\theta_{n-1}) - f(\theta^*)\right] - \gamma_n\mathbb{E}\|f'(\theta_{n-1})\|^2 + \gamma_n^2 L\sigma^2.$$

Using $f(\theta) - f(\theta^*) \leqslant \langle f'(\theta), \theta - \theta^*\rangle$ (from the convexity of f), we get, from Hölder's inequality:

$$\begin{aligned}
\Delta_{n-1}^2 = \Big[\mathbb{E}\left[f(\theta_{n-1}) - f(\theta^*)\right]\Big]^2 & \\
\leqslant \mathbb{E}\|f'(\theta_{n-1})\|^2 \times \mathbb{E}\|\theta_{n-1} - \theta^*\|^2 &\leqslant D_n\mathbb{E}\|f'(\theta_{n-1})\|^2 \qquad (42)
\end{aligned}$$

with Δ_n given in (37). We thus get our main recursion:

$$\Delta_n \leqslant (1+2\gamma_n^2L^2)\Delta_{n-1} - \frac{\gamma_n}{D_n}\Delta_{n-1}^2 + \gamma_n^2 L\sigma^2. \qquad (43)$$

Case: $\gamma_n = Cn^{-\alpha}$**,** $\alpha \in (1/2, 1)$**.** In this case, we have $D_n \leqslant D_\infty$. We make the following steps:

- We first show that the inequality in Eq. (43) may be replaced with an equality, which leads to a discrete time-difference equation which is analogous to the continuous-time Riccati ordinary differential equation (see, e.g., [27]).
- The behavior of the recursion changes depending on the sign of $-\frac{\gamma_n}{D_\infty}\Delta_{n-1}^2 + \gamma_n^2 L\sigma^2$, i.e., whether $\widetilde{\Delta}_{n-1}^2 \leqslant L\gamma_n\sigma^2 D_\infty$ or not. We will show that if this inequality is strict, then it remains strict and we can derive a simple bound on Δ_n.

Replacing inequalities by equalities. For all $n \in \mathbb{N}^*$, consider the recursion deduced from (43) by replacing inequality with equality:

$$\widetilde{\Delta}_n = (1 + 2\gamma_n^2 L^2)\widetilde{\Delta}_{n-1} - \frac{\gamma_n}{D_\infty}\widetilde{\Delta}_{n-1}^2 + \gamma_n^2 L\sigma^2, \tag{44}$$

with initial value $\widetilde{\Delta}_0 = \Delta_0$. It is worthwhile to note that, for any $n \in \mathbb{N}$,

$$\Delta_n \geqslant \widetilde{\Delta}_n\,.$$

This is due to the fact, that both sequences are upper-bounded by $(L/2)\,D_\infty$ (using the reasoning in Eq. (39)), where $D_\infty = (\delta_0 + \frac{\sigma^2}{L^2})\exp(\frac{2L^2C^2}{2\alpha-1})$, and that the function

$$t \mapsto \varphi_n(t) \stackrel{\text{def}}{=} (1 + 2\gamma_n^2 L^2)t - \frac{\gamma_n}{D_\infty}t^2\,, \tag{45}$$

is increasing on $[0, \frac{1+2\gamma_n^2L^2}{2\gamma_n}D_\infty] \supset [0, \frac{L}{2}D_\infty]$.

Relationship between $\widetilde{\Delta}_{n-1}^2$ **and** $L\gamma_n\sigma^2 D_\infty$**.** Denote

$$\varepsilon_n = (4L^{1/2}\sigma C^{3/2})^{-1} D_\infty^{1/2}\min\{1, n^{3\alpha/2-1}\}\,. \tag{46}$$

Since $\gamma_n^{1/2} - \gamma_{n+1}^{1/2} \geqslant \frac{C^{1/2}}{4n^{\alpha/2}}$ and (ε_n) is decreasing, we have for all $n > 0$,

$$\begin{aligned}
\gamma_n^{1/2}(1+\varepsilon_n)^{1/2} - \gamma_{n+1}^{1/2}(1+\varepsilon_{n+1})^{1/2} &\geqslant \gamma_n^{1/2}(1+\varepsilon_{n+1})^{1/2} - \gamma_{n+1}^{1/2}(1+\varepsilon_{n+1})^{1/2},\\
&\geqslant (\frac{C^{1/2}}{4n^{\alpha/2}})(1+\varepsilon_{n+1})^{1/2} \geqslant \frac{C^{1/2}}{4n^{\alpha/2}},\\
\varepsilon_n L^{1/2}\sigma\gamma_n^2 D_\infty^{-1/2} &= \sigma\gamma_n^2 D_\infty^{-1/2}(4L^{1/2}\sigma C^{3/2})^{-1} D_\infty^{1/2}\min\{1, n^{3\alpha/2-1}\},\\
&= \frac{C^{1/2}\min\{1, n^{3\alpha/2-1}\}}{4n^{2\alpha}} \leqslant \frac{C^{1/2}}{4n^{\alpha/2}},
\end{aligned}$$

leading to

$$\gamma_n^{1/2}(1+\varepsilon_n)^{1/2} - \gamma_{n+1}^{1/2}(1+\varepsilon_{n+1})^{1/2} \geqslant \varepsilon_n L^{1/2}\sigma\gamma_n^2 D_\infty^{-1/2}. \tag{47}$$

Let n_0 be the smallest n such that $\widetilde{\Delta}_{n-1}^2 \geqslant (1+\varepsilon_n)L\gamma_n\sigma^2 D_\infty$. Let us assume for now that n_0 is finite. We first establish that, for all $n \geq n_0$, we have $\widetilde{\Delta}_{n-1}^2 \geqslant (1+\varepsilon_n)L\gamma_n\sigma^2 D_\infty$. Indeed, if $\widetilde{\Delta}_{n-1} \geqslant (1+\varepsilon_n)^{1/2}L^{1/2}\gamma_n^{1/2}\sigma D_\infty^{1/2}$, then, since the function φ_n in Eq. (44) is increasing in $\widetilde{\Delta}_{n-1}$,

$$\begin{aligned}
\widetilde{\Delta}_n &\geqslant (1+2\gamma_n^2L^2)(1+\varepsilon_n)^{1/2}L^{1/2}\gamma_n^{1/2}\sigma D_\infty^{1/2} - \frac{\gamma_n}{D_\infty}((1+\varepsilon_n)^{1/2}L^{1/2}\gamma_n^{1/2}\sigma D_\infty^{1/2})^2 + \gamma_n^2 L\sigma^2\\
&\geqslant (1+\varepsilon_n)^{1/2}L^{1/2}\gamma_{n+1}^{1/2}\sigma D_\infty^{1/2} - \varepsilon\gamma_n^2 L\sigma^2 + L^{1/2}\sigma D_\infty^{1/2}(\gamma_n^{1/2}(1+\varepsilon_n)^{1/2} - \gamma_{n+1}^{1/2}(1+\varepsilon_{n+1})^{1/2})\\
&\geqslant (1+\varepsilon_{n+1})^{1/2}L^{1/2}\gamma_{n+1}^{1/2}\sigma D_\infty^{1/2},
\end{aligned}$$

because of our assumption regarding γ_n and ε_n (that led to Eq. (47)).

Thus for $n \geqslant n_0$, we have:

$$\begin{aligned}\widetilde{\Delta}_n &\leqslant (1+2\gamma_n^2L^2)\widetilde{\Delta}_{n-1} - \frac{\gamma_n}{D_\infty}\widetilde{\Delta}_{n-1}^2(1-(1+\varepsilon_n)^{-1}) \\ &= (1+2\gamma_n^2L^2)\widetilde{\Delta}_{n-1} - \frac{\gamma_n}{D_\infty}\widetilde{\Delta}_{n-1}^2\frac{\varepsilon_n}{1+\varepsilon_n}.\end{aligned}$$

If we denote $v_n = \widetilde{\Delta}_n \prod_{k=n_0+1}^n (1+2\gamma_k^2L^2)^{-1}$, for $n \geqslant n_0$, and $v_{n_0} = \widetilde{\Delta}_{n_0}$, then we have the inequality

$$\begin{aligned}v_n &\leqslant v_{n-1} - \frac{\gamma_n}{D_\infty}\frac{v_{n-1}^2}{(1+2\gamma_n^2L^2)}\prod_{k=n_0+1}^{n-1}(1+2\gamma_k^2L^2)\frac{\varepsilon_n}{1+\varepsilon_n} \\ &\leqslant v_{n-1} - \frac{\varepsilon_n}{1+\varepsilon_n}\frac{\gamma_n}{D_\infty}v_{n-1}^2\frac{1+2\gamma_{n-1}^2L^2}{1+2\gamma_n^2L^2} \\ &\leqslant v_{n-1} - \frac{\gamma_n}{2D_\infty}\frac{\varepsilon_n}{1+\varepsilon_n}v_{n-1}^2.\end{aligned}$$

We can now follow the standard argument from [20], i.e., divide by v_nv_{n-1} and obtain for $n \geqslant n_0$:

$$v_{n-1}^{-1} \leqslant v_n^{-1} - \frac{\gamma_n}{2D_\infty}\frac{\varepsilon_n}{1+\varepsilon_n}\frac{v_{n-1}}{v_n} \leqslant v_n^{-1} - \frac{\gamma_n}{2D_\infty}\frac{\varepsilon_n}{1+\varepsilon_n},$$

which leads to, by summing $n-n_0$ times,

$$v_{n_0}^{-1} \leqslant v_n^{-1} - \frac{1}{2D_\infty}\sum_{k=n_0+1}^n \frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k,$$

i.e., for $n \geqslant n_0$ (using the definition of n_0),

$$v_n \leqslant \frac{1}{\frac{1}{2D_\infty}\sum_{k=n_0+1}^n \frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k + v_{n_0}^{-1}} \leqslant \frac{1}{\frac{1}{2D_\infty}\sum_{k=n_0+1}^n \frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k + (1+\varepsilon_{n_0})^{-1/2}L^{-1/2}\gamma_{n_0}^{-1/2}\sigma^{-1}D_\infty^{-1/2}}$$

By assumption regarding the value of ε_n, for all n, we have

$$\begin{aligned}\gamma_n^{-1/2}(1+\varepsilon_n)^{-1/2} &= \sum_{k=1}^{n-1}(\gamma_{k+1}^{-1/2}(1+\varepsilon_{k+1})^{-1/2} - \gamma_k^{-1/2}(1+\varepsilon_k)^{-1/2}) + \gamma_1^{-1/2}(1+\varepsilon_1)^{-1/2} \\ &= \sum_{k=1}^{n-1}\frac{\gamma_k^{1/2}(1+\varepsilon_k)^{1/2} - \gamma_{k+1}^{1/2}(1+\varepsilon_{k+1})^{1/2}}{\gamma_k^{1/2}(1+\varepsilon_k)^{1/2}\gamma_{k+1}^{1/2}(1+\varepsilon_{k+1})^{1/2}} + \gamma_1^{-1/2}(1+\varepsilon_1)^{-1/2} \\ &\geqslant \sum_{k=1}^{n-1}(1+\varepsilon_{k+1})^{-1}\gamma_k^{-1/2}\gamma_{k+1}^{-1/2}(\gamma_k^{1/2} - \gamma_{k+1}^{1/2}) + \gamma_1^{-1/2}(1+\varepsilon_1)^{-1/2} \\ &\geqslant \frac{1}{2}\sum_{k=1}^n \gamma_k\frac{\varepsilon_k}{1+\varepsilon_k}L^{1/2}\sigma D_\infty^{-1/2}. \qquad (48)\end{aligned}$$

Thus, for $n \geqslant n_0$,

$$v_n \leqslant \frac{1}{\frac{1}{2D_\infty}\sum_{k=1}^n \frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k},$$

which leads to $\Delta_n \leqslant \frac{1}{\frac{1}{2D_\infty}\sum_{k=1}^n \frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k}\exp\left(2L^2\sum_{k=1}^n\gamma_k^2\right)$. Given Eq. (48), this is also true for $n < n_0$, and thus also true if n_0 is infinite. We thus get the desired bound because $D_\infty \geqslant \sigma^2/L$:

– For $\alpha > 2/3$,

$$\begin{aligned}\sum_{k=1}^{n} \frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k &\geqslant (1+4L^{1/2}\sigma C^{3/2}D_\infty^{-1/2})^{-1} C\varphi_{1-\alpha}(n) \\ &\geqslant (1+4L^{3/2}C^{3/2})^{-1} C\varphi_{1-\alpha}(n).\end{aligned}$$

– For $\alpha \leqslant 2/3$,

$$\begin{aligned}\sum_{k=1}^{n} \frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k &= \sum_{k=1}^{n} \frac{1}{1+\varepsilon_k^{-1}}\gamma_k \\ &= \sum_{k=1}^{n} \frac{1}{1+4L^{1/2}\sigma C^{3/2}D_\infty^{-1/2}k^{1-3\alpha/2}} Ck^{-\alpha} \\ &\geqslant (1+4L^{1/2}\sigma C^{3/2}D_\infty^{-1/2})^{-1} C\varphi_{\alpha/2}(n) \\ &\geqslant (1+4L^{3/2}C^{3/2})^{-1} C\varphi_{\alpha/2}(n).\end{aligned}$$

Study of recursion for $\gamma_n = Cn^{-1/2}$. For all $\alpha \in (1/2, 1]$, we can replace D_∞ by D_n in the final bound. Note that this may not lead to the best possible bound; we thus get the desired result, which is also valid for $\alpha = 1/2$. ■

D.1 Proof for Theorem 5

Proof The proof technique is the same as in previous theorems.

Derivation of deterministic recursion. Since $\|f_n'(\theta)\| \leqslant B$ almost surely, then,

$$\delta_n \leqslant \delta_{n-1} + B^2\gamma_n^2, \tag{49}$$

leading to a bound $\delta_n \leqslant D_n \stackrel{\text{def}}{=} \delta_0 + B^2 \sum_{k=1}^{n} \gamma_k^2$. Using (40), we obtain

$$\mathbb{E}\left[f(\theta_n) | \mathcal{F}_{n-1} \right] \leqslant f(\theta_{n-1}) - \gamma_n \|f'(\theta_{n-1})\|^2 + \frac{\gamma_n^2 LB^2}{2},$$

which implies, together with (42), the following recursion (which replaces (43)),

$$\Delta_n \leqslant \Delta_{n-1} - \frac{\gamma_n}{D_n}\Delta_{n-1}^2 + \frac{1}{2}\gamma_n^2 LB^2, \tag{50}$$

where Δ_n is defined in (37).

Study for $\alpha > 1/2$. We can then follow exactly the same proof than for Theorem 4 for $\alpha \in [1/2, 1]$, and obtain the desired result for $\alpha > 1/2$ (note that the absence of the multiplicative factor $(1 + 2L^2\gamma_n^2)$ makes the problem a little easier).

Study for $\alpha < 1/2$. For $\alpha < 1/2$, we consider

$$\varepsilon_n = (4L^{1/2}BC^{3/2})^{-1} D_n^{1/2} n^{3\alpha/2-1} \geqslant (4L^{1/2}C^{1/2})^{-1} \varphi_{1-2\alpha}(n)^{1/2} n^{3\alpha/2-1},$$

which is a decreasing sequence. We can then follow the same reasoning than for the proof of Theorem 4.

The recursion defining $\widetilde{\Delta}_n$ provides an upper bound to Δ_n as soon as $D_n/2\gamma_n \geqslant L/2$, i.e., $\gamma_n \leqslant 1/L$, which simply implies that the reasoning from Theorem 4 can only be applied for n large enough.

Following the same reasoning than for Theorem 4, we get, using $D_k \leqslant (\delta_0 + B^2C^2)\varphi_{1-2\alpha}(k)$ and $\varepsilon_n \leqslant (4L^{1/2}BC^{3/2})^{-1}(\delta_0 + B^2C^2)^{1/2}$:

$$\begin{aligned}
\Delta_n &\leqslant \frac{1}{\frac{1}{2}\sum_{k=1}^n \frac{1}{D_k}\frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k} \\
&= \frac{1}{\frac{1}{2}\sum_{k=1}^n \frac{1}{D_k}\frac{1}{1+\varepsilon_k^{-1}}\gamma_k} \\
&= \frac{1}{\frac{1}{2}\sum_{k=1}^n \frac{1}{D_k}\frac{1}{1+4L^{1/2}BC^{3/2}D_k^{-1/2}k^{1-3\alpha/2}}Ck^{-\alpha}} \\
&\leqslant \frac{1}{\frac{1}{2}\sum_{k=1}^n \frac{1}{D_k}\frac{1}{(1+4L^{1/2}BC^{3/2})D_k^{-1/2}k^{1-3\alpha/2}}Ck^{-\alpha}} \\
&= \frac{2(1+4L^{1/2}BC^{3/2})/C}{\sum_{k=1}^n \frac{1}{D_k^{1/2}}k^{\alpha/2-1}}
\end{aligned}$$

Therefore

$$\begin{aligned}
\Delta_n &\leqslant \frac{2(1+4L^{1/2}BC^{3/2})(\delta_0 + B^2C^2)^{1/2}/C}{\sum_{k=1}^n \frac{1}{\varphi_{1-2\alpha}(k)^{1/2}}k^{\alpha/2-1}} \\
&\leqslant \frac{2(1+4L^{1/2}BC^{3/2})(\delta_0 + B^2C^2)^{1/2}/C}{\sum_{k=1}^n \frac{(1-2\alpha)^{1/2}}{k^{1/2-\alpha}}k^{\alpha/2-1}} \\
&\leqslant \frac{2(1+4L^{1/2}BC^{3/2})(\delta_0 + B^2C^2)^{1/2}/C}{(1-2\alpha)^{1/2}\varphi_{3\alpha/2-1/2}(n)} \\
&\leqslant \frac{2(1+4L^{1/2}BC^{3/2})(\delta_0 + B^2C^2)^{1/2}/C}{(1-2\alpha)^{1/2}\varphi_{3\alpha/2-1/2}(n)},
\end{aligned}$$

which leads to the desired result for $\alpha < 1/2$. ■

D.2 Proof of Theorem 6

Proof We follow the proof of [8, 7] by adapting it to the smooth case. Note first that by convexity

$$f\left(n^{-1}\sum_{k=0}^{n-1}\theta_k\right) \leqslant \frac{1}{n}\sum_{k=0}^{n-1} f(\theta_k)\,.$$

Since (still by convexity of f) $f(\theta^*) \geq f(\theta_{k-1}) + \langle f'(\theta_{k-1}), \theta^* - \theta_{k-1}\rangle$, the previous inequality implies

$$f(\theta_{k-1}) - f(\theta^*) \leqslant \langle f'(\theta_{k-1}), \theta_{k-1} - \theta^*\rangle\,. \tag{51}$$

Moreover, we have:

$$\begin{aligned}
\mathbb{E}\left[\|\theta_k - \theta^*\|^2 \middle| \mathcal{F}_{k-1}\right] \leqslant \|\theta_{k-1} - \theta^*\|^2 - 2\gamma_k\langle f'(\theta_{k-1}), \theta_{k-1} - \theta^*\rangle \\
+ \gamma_k^2\mathbb{E}(\|f_k'(\theta_{k-1})\|^2|\mathcal{F}_{k-1})\,;
\end{aligned} \tag{52}$$

The inequality [20, Eq. (2.1.8)] yields

$$L^{-1} \left\| f_k'(\theta_{k-1}) - f_k'(\theta^*) \right\|^2 \leqslant \langle f_k'(\theta_{k-1}) - f_k'(\theta^*), \theta_{k-1} - \theta^* \rangle \ ,$$

which implies that

$$\left\| f_k'(\theta_{k-1}) \right\|^2 \leq 2 \ \left\| f_k'(\theta^*) \right\|^2 + 2L \left\langle f_k'(\theta_{k-1}) - f_k'(\theta^*), \theta_{k-1} - \theta^* \right\rangle \ .$$

Plugging this inequality in Eq. (52) then implies

$$\delta_k \leqslant \delta_{k-1} - 2\gamma_k \left(1 - \gamma_k L\right) \mathbb{E}\left[\langle f'(\theta_{k-1}), \theta_{k-1} - \theta^* \rangle\right] + 2\gamma_k^2 \sigma^2 \, ,$$

where $\delta_k = \mathbb{E}\left[\left\| \theta_k - \theta^* \right\|^2 \right]$, showing that

$$2\gamma_k \left[1 - \gamma_k L\right] \mathbb{E}\left[\langle f'(\theta_{k-1}), \theta_{k-1} - \theta^* \rangle\right] \leqslant \delta_{k-1} - \delta_k + 2\gamma_k^2 \sigma^2 \, . \tag{53}$$

Define $n_0 = \inf \left\{ k \in \mathbb{N}, (1 - \gamma_k L) \geq 1/2 \right\}$. For any $k \geq n_0$, $1 - \gamma_k L \leq 1/2$ and therefore,

$$\mathbb{E}\left[\langle f'(\theta_{k-1}), \theta_{k-1} - \theta^* \rangle\right] \leq \gamma_k^{-1} \left(\delta_{k-1} - \delta_k + 2\gamma_k^2 \sigma^2 \right) \ .$$

Note that by integrating by parts,

$$\sum_{k=n_0+1}^{n} \gamma_k^{-1} \left[\delta_{k-1} - \delta_k\right] = \gamma_{n_0+1}^{-1} \delta_{n_0} + \sum_{k=n_0+1}^{n-1} \delta_k \left[\gamma_{k+1}^{-1} - \gamma_k^{-1}\right] - \gamma_n^{-1} \delta_n \, .$$

Using that $\delta_n \leqslant D_n$, where D_n is defined in (39), and (D_n) is non-decreasing, the previous identity shows that

$$\sum_{k=n_0+1}^{n} \gamma_k^{-1} \left[\delta_{k-1} - \delta_k\right] \leqslant D_n \gamma_n^{-1} \, . \tag{54}$$

Combining Eq. (53) and Eq. (54) shows that For $k \in \{1, \ldots, n_0\}$, under **(H2')**,

$$\left| f(\theta_k) - f(\theta^*) \right| \leq \left| \int_0^1 \langle f'(\theta^* + t(\theta_k - \theta^*)) - f'(\theta^*), \theta_k - \theta^* \rangle \, \mathrm{d}t \right| \leq L/2 \left\| \theta_k - \theta^* \right\|^2 \ .$$

Combining these two inequalities finally yields to

$$\mathbb{E}\left[f\left(n^{-1} \sum_{k=0}^{n-1} \theta_k \right) \right] - f(\theta^*) \leqslant \frac{1}{n} \Big[\frac{D_n}{\gamma_n} + \frac{1}{2} \sigma^2 \sum_{k=n_0+1}^{n} \gamma_k + \frac{L}{2} \sum_{k=1}^{n_0} D_k \Big] \, ,$$

where we used the Lipschitz-continuity of f.

Case $\gamma_n = C n^{-1/2}$.

In this case, for $L^2 C^2 < 1/4$, we have $D_n = \left(\delta_0 + \frac{\sigma^2}{L^2} \right) n^{2L^2C^2}$. This leads to an upper bound of the form

$$\frac{1}{n^{1/2 - 2L^2C^2}} \frac{1}{C} \left(\delta_0 + \frac{\sigma^2}{L^2} \right) + \frac{C\sigma^2}{n} (n^{-1/2} - n_0^{-1/2}) + \frac{1}{n} \frac{L}{2} \left(\delta_0 + \frac{\sigma^2}{L^2} \right) \frac{1}{2L^2C^2 + 1} n_0^{2L^2C^2 + 1}$$

leading to, for $n_0 = (2LC)^2$,

$$\frac{1}{n^{1/2 - 2L^2C^2}} \frac{1}{C} \left(\delta_0 + \frac{\sigma^2}{L^2} \right) + \frac{\sigma^2}{n^{1/2}} + \frac{L}{2} \left(\delta_0 + \frac{C\sigma^2}{L^2} \right) \frac{1}{2L^2C^2 + 1} (4L^2C^2)^{2L^2C^2 + 1} \frac{1}{n}$$

Case $\gamma_n = Cn^{-\alpha}$, $\alpha \in (1/2, 1)$. In this case, we have $D_n = \left(\delta_0 + \frac{\sigma^2}{L^2}\right) \exp\left(\frac{2L^2C^2}{2\alpha-1}\right)$, leading to the upper bound

$$\left(\delta_0 + \frac{\sigma^2}{L^2}\right) \exp\left(\frac{2L^2C^2}{2\alpha-1}\right) \frac{1}{C} n^{\alpha-1} + \frac{C\sigma^2}{n} \varphi_{1-\alpha}(n) + \frac{L}{2} \left(\delta_0 + \frac{\sigma^2}{L^2}\right) \exp\left(\frac{2L^2C^2}{2\alpha-1}\right) \frac{(2LC)^{1/\alpha}}{n}$$

Case $\gamma_n = Cn^{-1}$. In this case, we have $D_n = \left(\delta_0 + \frac{C\sigma^2}{L^2}\right) \exp\left(\frac{L^2C^2\pi^2}{6}\right)$, leading to the upper bound

$$\left(\delta_0 + \frac{C\sigma^2}{L^2}\right) \exp\left(\frac{L^2C^2\pi^2}{6}\right) \frac{1}{C} + \frac{C\sigma^2}{2} \frac{\ln(n)}{n} + \frac{L}{2} \left(\delta_0 + \frac{\sigma^2}{L^2}\right) \exp\left(\frac{L^2C^2\pi^2}{6}\right) .$$

We can summarize the results as follows:

$$\Delta_n \leqslant \left(\delta_0 + \frac{\sigma^2}{L^2}\right) \exp\left(2L^2C^2\varphi_{1-2\alpha}(n)\right) \left[\frac{1}{C} n^{\alpha-1} + \frac{L}{2} \frac{(2LC)^{1/\alpha}}{n}\right] + \frac{\sigma^2}{2n} C\varphi_{1-\alpha}(n) \tag{55}$$

and further bound $\left[\frac{1}{C} n^{\alpha-1} + \frac{L}{2} \frac{(2LC)^{1/\alpha}}{n}\right]$ by $\frac{1}{C}\left[1 + (2LC)^{1+\frac{1}{\alpha}}\right]$, which leads to the desired result. ■

E Proof of Theorem 7

Proof We follow [7, 15, 14]. Since $\mathbb{E}(\|f_k'(\theta_{k-1})\|^2 | \mathcal{F}_{k-1}) \leq B^2$ almost-surely, Eq. (52) rewrites

$$\mathbb{E}\left[\|\theta_k - \theta^*\|^2 \middle| \mathcal{F}_{k-1}\right] \leqslant \|\theta_{k-1} - \theta^*\|^2 - 2\gamma_k \langle f'(\theta_{k-1}), \theta_{k-1} - \theta^* \rangle + \gamma_k^2 B^2 . \tag{56}$$

which implies that

$$2\gamma_k \mathbb{E}\left[\langle f'(\theta_{k-1}), \theta_{k-1} - \theta^* \rangle\right] \leqslant \delta_{k-1} - \delta_k + \gamma_k^2 B^2 . \tag{57}$$

Since, on the other hand, for any $n \in \mathbb{N}$, $\delta_n \leq D_n = \delta_0 + B^2 \sum_{k=1}^n \gamma_k^2$, (51) shows that

$$\mathbb{E}\left[f\left(n^{-1} \sum_{k=0}^{n-1} \theta_k\right)\right] - f(\theta^*) \leqslant \frac{1}{2n}\left[\frac{D_n}{\gamma_n} + B^2 \sum_{k=0}^{n-1} \gamma_k\right] .$$

This leads to the bound

$$\frac{n^{\alpha-1}}{2C}(\delta_0 + C^2B^2\varphi_{1-2\alpha}(n)) + \frac{B^2}{2n}\varphi_{1-\alpha}(n).$$

■

F Additional experiments

Medium-scale experiments with linear logistic regression. We consider two situations where $\mathcal{H} = \mathbb{R}^p$: (a) the "alpha" dataset from the Pascal large scale learning challenge (`http://largescale.ml.tu-berlin.de/`), for which $p = 500$ and $n = 50000$, and (b) a synthetic example where $p = 100$, $n = 100000$, we generate the input data i.i.d. from a multivariate Gaussian distribution with mean

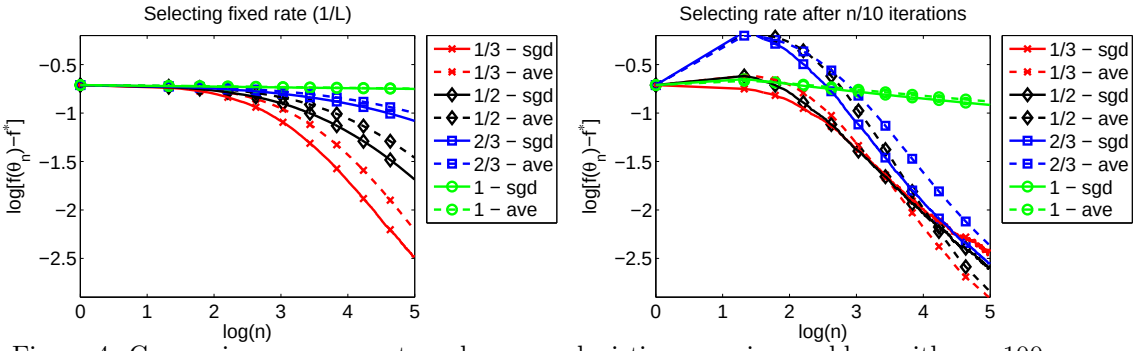

Figure 4: Comparison on a non strongly convex logistic regression problem with $p = 100$.

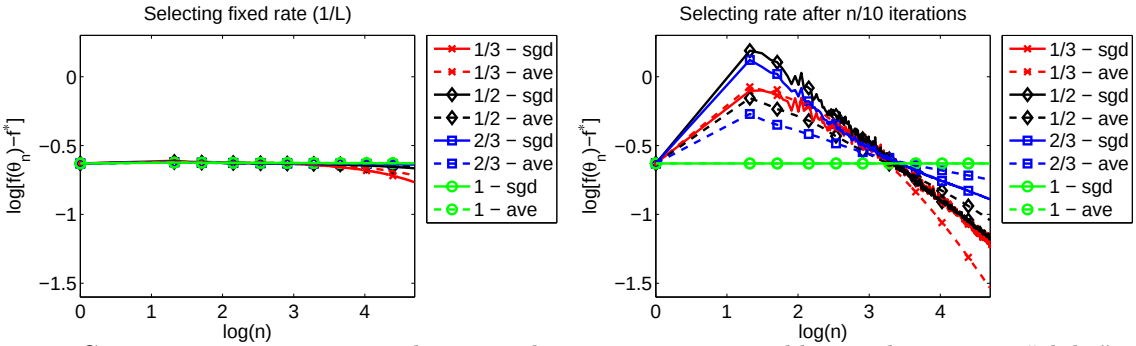

Figure 5: Comparison on a non strongly convex logistic regression problem with $p = 100$. "alpha" dataset.

zero and a covariance matrix sampled from a Wishart distribution with p degrees of freedom (thus with potentially bad condition number), and the output is obtained through a classification by a random hyperplane. In Figure 4 and Figure 5, we try two ways of selecting the constant C for $\gamma_n = Cn^{-\alpha}$: (1) a fixed rate equal to $1/L$ suggested by our analysis to avoid large constants (left plots), for which the convergence speed may be too slow, hinting at the fact that our global bounds involving the Lipschitz constants may be locally far too pessimistic; (2) an adaptive way where we consider the lowest test error after $n/10$ iterations (right plots), for which convergence is much faster, hinting at the fact that a truly adaptive sequence (γ_n) instead of a fixed one is clearly an important avenue for future research.

References

[1] M. N. Broadie, D. M. Cicek, and A. Zeevi. General bounds and finite-time improvement for stochastic approximation algorithms. Technical report, Columbia University, 2009.

[2] H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer-Verlag, second edition, 2003.

[3] O. Yu. Kul′chitskiĭ and A. È. Mozgovoĭ. An estimate for the rate of convergence of recurrent robust identification algorithms. *Kibernet. i Vychisl. Tekhn.*, 89:36–39, 1991.

[4] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[5] D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report 781, Cornell University Operations Research and Industrial Engineering, 1988.

[6] V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

[7] Y. Nesterov and J. P. Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6):1559–1568, 2008.

[8] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[9] L. Bottou and Y. Le Cun. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005.

[10] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS), 20*, 2008.

[11] S. Shalev-Shwartz and N. Srebro. SVM optimization: inverse dependence on training set size. In *Proc. ICML*, 2008.

[12] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proc. ICML*, 2007.

[13] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Conference on Learning Theory (COLT)*, 2009.

[14] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 9:2543–2596, 2010.

[15] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, 2009.

[16] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2006.

[17] R. Durrett. *Probability: theory and examples*. Duxbury Press, third edition, 2004.

[18] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2001.

[19] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[20] Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Kluwer Academic Publishers, 2004.

[21] K. Sridharan, N. Srebro, and S. Shalev-Shwartz. Fast rates for regularized objectives. *Advances in Neural Information Processing Systems*, 22, 2008.

[22] N. N. Vakhania, V. I. Tarieladze, and S. A. Chobanyan. *Probability distributions on Banach spaces*. Reidel, 1987.

[23] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization, 2010. Tech. report, Arxiv 1009.0571.

[24] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[25] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, 2001.

[26] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002.

[27] C. D. Ahlbrandt and A. C. Peterson. *Discrete Hamiltonian systems: Difference equations, continued fractions, and Riccati equations*, volume 16. Kluwer Academic Pub., 1996.